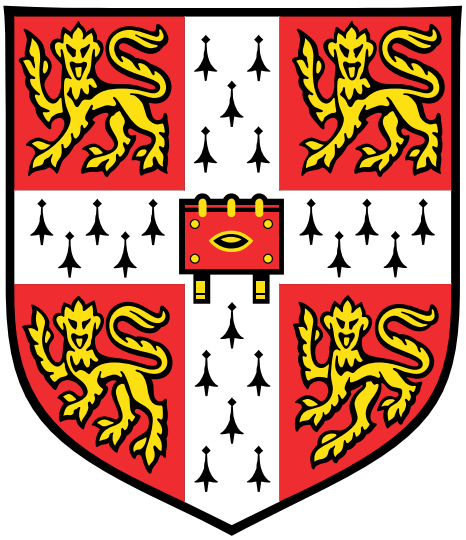
Statistical co-analysis of high dimensional association studies



James Liley

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College

September 2017

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words including footnotes, tables and equations, and excluding appendices and bibliography, and has fewer than 150 figures.

James Liley
September 2017

Acknowledgements

First and foremost I would like to thank my primary supervisor, Chris Wallace. Chris's commitment to statistical integrity, insight and knowledge, support, patience, and willingness to debate my more far-fetched ideas have made my PhD an extraordinary experience, during which I have learned a great deal. Thanks also go to my former primary supervisor John Todd, and secondary supervisor Eoin Mckinney, for their support, useful comments, and alternative angles on my work. I thank Andrew Morris and Oliver Stegle for their careful reading of this thesis, and for a useful discussion during the viva examination.

My thesis and time in the UK was funded by the NIHR Biomedical Research Council, without whose generous support I would have been unable to undertake this PhD. I studied on the Mathematical Genomics and Medicine PhD programme, which is funded by the Wellcome Trust. I thank both of these organisations not just for supporting me, but for their immense role in medical research and commitment to global health.

All the work in this thesis involves analysis of data collected by others, and I extend my thanks to all the groups I have collaborated with. These include the University of Manchester, both for access to their data on JIA and for very helpful discussions; the Smith group in the Department of Medicine for the privilege of collaborating on their EGPA work, and the former Diabetes and Inflammation Laboratory, University of Cambridge for accommodation during the first part of my PhD and for datasets relating to T1D.

I thank my wonderful recent office-mates Oliver Burren, Mary Fortune and Stasia Grinberg both formally for helping me out and for the useful discussions and views on the field we work in, and informally for being great friends and for making a working environment I looked forward to coming in to every day. My thanks also go to former labmates Nick Cooper, Chris Eijsbouts and Niko Pontikos, who did the same earlier in my PhD, and the other members of the Wallace group for their insight and help.

Outside of work, I extend my thanks to the University of Cambridge Squash club and Cambridgeshire Squash (and our counterparts at Oxford for keeping us humble), and the Cambridge University Music society, particularly the Cambridge University Chamber Orchestra. Although at times more nerve-racking than my PhD, the times I had with both

will be long treasured. I also thank Darwin College and the DCSA and all the friends I made there and elsewhere. Finally, I thank my parents, my siblings, and Eleanor for their unwavering support and grounding.

Specific acknowledgements for each chapter are listed below.

Chapter 2. I thank John Todd for help in interpreting the results, and Cathryn Lewis and Steve Eyre for the list of SNPs used for genomic control. Acknowledgements regarding data collection for this chapter can be found in [Liley and Wallace, 2015]. In particular I thank all volunteers for their support and participation in this study.

Chapter 3. I acknowledge the same sources of samples as for chapter 2. Acknowledgements for JIA samples can be found in [Hinks et al., 2013]. I thank co-authors as listed in [Lyons et al., 2017] for sample collection, quality control, imputation and conducting the main GWAS in the work on EGPA.

Chapter 4. I thank my supervisor Chris Wallace, and colleague Jenn Asimit for helpful comments and review of this work.

Chapter 5. I acknowledge the help of the Diabetes and Inflammation Laboratory Data Service for access and quality control procedures on the datasets used in this study.

Chapter 6. I thank my collaborators in Manchester (Wendy Thomson, Steve Eyre, Anne Hinks, John Bowes, Jo Cobb Sam Smith and Sunil Sampath) for their sample recruitment, generation and quality control of data, and for useful discussion leading to the development of the methods described in the chapter.

General. All work except that in chapters 3 and 6 re-analyses previously published datasets. All patient data were handled in accordance with the policies and procedures of the participating organisations. Computation was performed either on the Diabetes and Inflammation Laboratory computing cluster or the Cambridge high-performance computing cluster (HPC). Analyses were conducted in R and Mathematica.

All work was on human samples, many from people with serious disease. My final thanks correspondingly goes to all patients who provided samples for these analyses, and for all those who contribute to the development of medical science.

Abstract

Modern medical practice and science involve complex phenotypic definitions. Understanding patterns of association across this range of phenotypes requires co-analysis of high-dimensional association studies in order to characterise shared and distinct elements. In this thesis I address several problems in this area, with a general linking aim of making more efficient use of available data. The main application of these methods is in the analysis of genome-wide association studies (GWAS) and similar studies.

Firstly, I developed methodology for a Bayesian conditional false discovery rate (cFDR) for levering GWAS results using summary statistics from a related disease. I extended an existing method to enable a shared control design, increasing power and applicability, and developed an approximate bound on false-discovery rate (FDR) for the procedure. Using the new method I identified several new variant-disease associations. I then developed a second application of shared control design in the context of study replication, enabling improvement in power at the cost of changing the spectrum of sensitivity to systematic errors in study cohorts. This has application in studies on rare diseases or in between-case analyses.

I then developed a method for partially characterising heterogeneity within a disease by modelling the bivariate distribution of case-control and within-case effect sizes. Using an adaptation of a likelihood-ratio test, this allows an assessment to be made of whether disease heterogeneity corresponds to differences in disease pathology. I applied this method to a range of simulated and real datasets, enabling insight into the cause of heterogeneity in autoantibody positivity in type 1 diabetes (T1D). Finally, I investigated the relation of subtypes of juvenile idiopathic arthritis (JIA) to adult diseases, using modified genetic risk scores and linear discriminants in a penalised regression framework.

The contribution of this thesis is in a range of methodological developments in the analysis of high-dimensional association study comparison. Methods such as these will have wide application in the analysis of GWAS and similar areas, particularly in the development of stratified medicine.

Table of contents

List of figures

List of tables

Nomenclature

Abbreviations

MAF	Minor allele frequency
AAV	Anti-neutrophil cytoplasmic antibody associated vasculitis
AF	Allele frequency
BVN	Bivariate normal distribution
CD/CRO	Crohn's disease
CR	Call rate
EC	Eosinophil count
EGPA	Eosinophilic granulomatosis with polyangiitis (formerly Churg-Strauss disease)
EO	Extended oligoarthritis
FDR	False-discovery rate
FWER	Family-wise error rate
GCTA	Genome-wide complex trait analysis
GRS	Genetic (polygenic) risk score
GWAS	Genome-wide association study
GWS	Genome-wide significance
HWE	Hardy-Weinberg equilibrium
IBD	Inflammatory bowel disease

ILAR	International League of Associations for Rheumatology
JIA	Juvenile Idiopathic Arthritis
LDSC	LD-Score regression
MAF	Minor allele frequency
MVN	Multivariate normal distribution
OR	Odds ratio
PCA	Principal Component Analysis
PC	Principal component
PO	Persistent oligoarthritis
QC	Quality control
RA	Rheumatoid arthritis
RF	Rheumatoid factor
ROC	Receiver-operator characteristic curve
SE	Standard error
SNP	Single nucleotide polymorphism
T1D	Type 1 diabetes
T2D	Type 2 diabetes
UC	Ulcerative colitis

Symbols

$\text{logit}(.)$	Logit function; $\text{logit}(x) = (1+e^{-x})^{-1}$
χ^2_n	χ^2 distribution with n degrees of freedom, NCP 0
$\text{erf}(z)$	error function
$\text{erfc}(z)$	$1-\text{erf}(z)$ (complementary error function)

r_g	Narrow - sense genetic correlation
ρ_g	Narrow - sense genetic covariance
H^2	Broad-sense heritability
h^2	Narrow-sense heritability
$\mathbf{1_S}$	Indicator function for set S
$\int_{\mathbb{R}} \dots$	$\int_{-\infty}^{\infty} \dots$
λ	genomic inflation factor [Devlin et al., 2001]
$\wedge/\cap$	AND (logical conjunction) or set union
$\vee/\cup$	OR (logical disjunction) or set intersection
m_i, m_j	observed (study) minor allele frequencies in groups i, j
I_N/I_{MN}	Identity matrix of dimension $N^2/M \times N$
$1_N/1_{MN}$	Matrix of 1's of dimension $N^2/M \times N$
μ_i, μ_j	population minor allele frequencies in groups i, j
$N_{\mu,\Sigma}(\mathbf{z})$	Density of (usually multivariate) Gaussian with mean μ, variance Σ at $\mathbf{z}$
$N_{\Sigma}(\mathbf{z})$	density of MVN with mean $\mathbf{0}$, covariance Σ at $\mathbf{z}$
n_i, n_j	sample sizes in groups i, j
$\Phi(z)$	$\int_{-\infty}^{z} N_{0,1}(x)dx$
p_s	p-value for study s
P_s	random variable associated with p-value for study s; context-dependent
$Q_X(x)$	Quantile function of distribution X at x
$X(\Omega)$	Image of a random variable X
$\{A : B\}$	Set of all A such that B holds
$U(a,b)$	Uniform distribution on $[a,b]$, or when $a = 0$, $b = 1$, on $(0,1]$

ζ_s	(generally) $E(Z_s)$
z_s	z score associated with p-value, signed or unsigned; $z_s = -\Phi^{-1}\left(\frac{p_s}{2}\right)$
Z_s	random variable associated with z_s; context dependent

Chapter 1

Introduction

1.1 Foreword

The scientific background to this thesis involves several scientific and mathematical disciplines, including genetics, pathology, probability theory and computer science. In this introduction I give an overview of the parts of these disciplines relevant to my project. Each chapter begins with a more specific introduction to the topics and literature it covers.

I begin with a discussion of the history of association testing in medicine, landmarked by the successive introduction of epidemiological methods and increase in the number of independent variables (dimensionality). As part of this, I discuss the motivation for conducting and comparing association studies, and the reasons behind the development of statistical methods for their analysis.

A major application and testing ground for my methods is the field of genome-wide association study (GWAS). I discuss the motivation for conducting these studies, and describe some of the relevant statistical procedures. I then categorise and discuss procedures for comparing association studies, and describe some heuristic arguments regarding efficiency of data use. Finally, I give an overview of the structure of the remainder of this thesis, and give an indication of the position of each chapter of this thesis in the wider field of statistical genetics.

The major contribution of my work has been the development of new statistical methods for joint analysis of GWAS and other association studies, with a focus on making the best use of input data. In general, the chapters of this thesis are independent linked experiments of similar prominence, rather than a single linear narrative in which one experiment directly leads to another. In the spirit of computational biology, I have represented this below as an

undirected graph, with vertices representing chapters and edges representing linking ideas (figure 1.1).

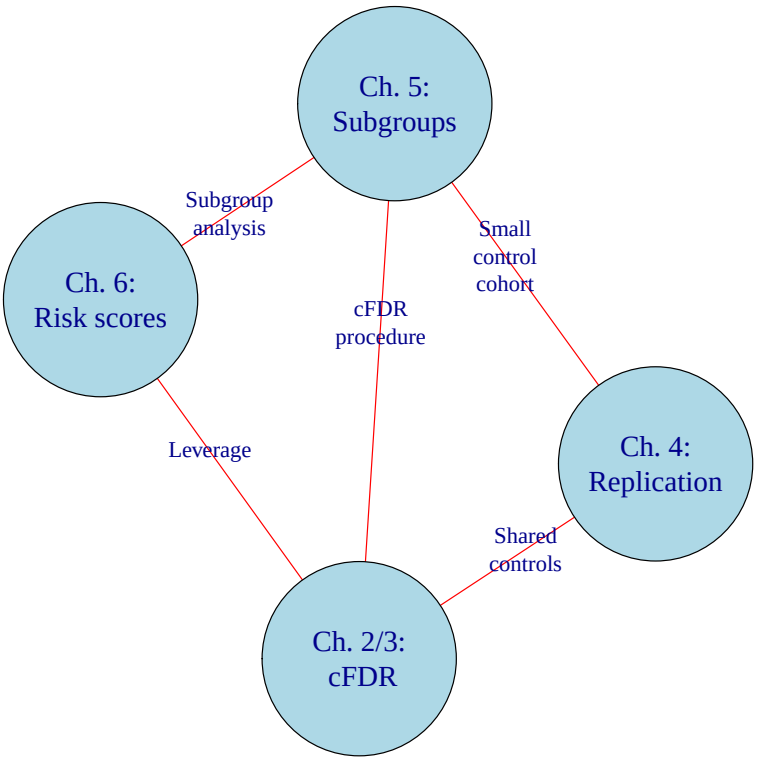


Fig. 1.1 Diagram of chapters of this thesis, with linking ideas.

1.2 Association testing and case-control studies

1.2.1 Philosophy

The question of whether two variables of interest are dependent - in practical terms, whether information about one can tell us something about the other - is the essence of many questions in quantitative science, particularly medicine. Establishing association between observable phenomena is the first step in most investigative processes, usually analysed prior to investigating more specific phenomena such as causality [Hill, 1965]. A range of statistical procedures can be used to test for independence, commonly in a frequentist framework in which the joint distribution of the variables in question can be predicted under a null hypothesis of independence.

This area of applied statistical theory is, in a sense, a formalisation of intuition. For much of the history of medicine, association analyses were interpreted 'by eye', with rigorous probabilistic procedures only established in the past several centuries [Hacking, 1990]. More precise analytic procedures enable more subtle associations to be found or reduce of the amount of data required to assert a result. An important analytic procedure devised for this purpose is the case-control study [Paneth et al., 2004].

A critical step in conducting a case/control study is the assembly of a representative group of patients with similar illnesses [Paneth et al., 2004]. In order to maximise the power of association studies, it is sensible to combine patients with only subtle differences in disease into general case cohorts, seeking to understand the common elements of the disease in question. The question of what patients and their illnesses have in common (and, by proxy, what makes them unwell) invites the parallel question of why their illnesses differ. The analysis of differences takes several forms, and includes searching for similarity between nominally distinct diseases as well as for heterogeneity in diseases historically considered homogenous. Studies of this type are becoming more feasible as phenotype characterisation improves and study sizes increase [Robinson et al., 2014]. Both the comparison of case-control studies on separate diseases and the analysis of heterogeneity within-diseases present new statistical challenges, some of which I consider in this thesis.

1.2.2 Historical development

Early association tests

The use of association testing to guide medical practice was widespread long before formalised procedures were developed, and evidence of some form of association analysis is found even in the earliest medical texts. Ancient records such as the Edwin Smith Papyrus [Feldman and Goodrich, 1999] and the Babylonian Sakikku (diagnostic handbook) [Heeßel, 2004] both detail collections of case histories with a view towards identifying commonalities in presentation. For example, Smith's papyrus, an incomplete copy of a document dating from c. 3000BC, demonstrates awareness of an association between traumatic spinal injury and limb paralysis. This knowledge effectively arose from a rudimentary observational study of injured patients.

The use of patient cohorts implied the expectation that patients with similar clinical conditions would show similar response to interventions, and the assumption that such cohorts were representative of wider populations with the disease. This was acknowledged by the tenth-century Persian scholar Avicenna, who, in a protocol for determining drug efficacy,

included the directive that an observed effect in a patient cohort must occur in repeated samples from a population to distinguish a potential effect from randomness [Shoja et al., 2011].

Formalisation of the hypothesis testing procedure took place relatively recently. The first use of an explicit 'null hypothesis' was probably by Laplace [Laplace, 1781], who assessed the difference in frequency ΔF of sexes amongst children born in London and Paris. By considering the distribution of ΔF under the null hypothesis H_0 that the population frequencies were equal, Laplace demonstrated that a deviance from equality as large as that observed was unlikely under H_0, and thus generated the first known example of a p-value:

$$p = Pr(|\Delta F| \geq \text{Observed}|H_0) \tag{1.1}$$

Development of the case-control study

The first case-control study is difficult to identify, with different elements of the procedure being introduced separately. Case-control studies are mainly concerned with the aetiology of diseases rather than their management, and hence diverge from Avicenna's and the ancient Egyptian's primary aims [Paneth et al., 2004]. Early studies were incentivised by epidemics, during which determining aetiology was of particular importance.

An early success of epidemiology was John Snow's investigation into causes of cholera in 1854. By considering the location of a group of cases, Snow determined that a likely source for the outbreak was a water pump, eventually leading to the understanding of the water-borne nature of cholera. Snow's study notably lacked a control set; however, a corresponding investigation by Reverend Henry Whitehead considered controls instead, validating Snow's findings and constructing a case/control odds ratio for exposure to the water pump [Paneth et al., 2004].

The first case-control study in the modern sense was probably Janet Lane-Claypon's 1920s study on breast cancer [Lane-Claypon et al., 1926]. In a study of 500 women from across England with a breast cancer history and 500 matched controls, Lane-Claypon identified several obstetric and gynaecological risk factors for the disease, many of which have held up to modern analyses. Even at this early stage in the field, the importance of controlling for potential biases was acknowledged and discussed [Paneth et al., 2004, Press, 2008].

The study was a landmark in many respects. It combined the existing approaches of case-series analysis with the procedure of anamnesis (focused and systematic history-taking) and systematic comparison of differences between groups [Paneth et al., 2004], and demonstrated the potential contribution such studies could make to public health. A subsequent replication

study in the USA [Wainwright, 1931] helped set a precedent for accurate description of research methods in the interests of subsequent replication [Paneth et al., 2004].

1.2.3 Statistical hypothesis testing

The modern procedure of frequentist hypothesis testing begins with identifying a 'null hypothesis' H_0, which may be rejected or not. In the context of association studies this hypothesis is typically of statistical independence between variables. An observed dataset is quantified with a univariate test statistic associated with a random variable X which has a known distribution under H_0, and the p-value of an observation x of X is then generically defined as

$$Pr(X \geq x|H_0) \tag{1.2}$$

where '$\geq$' is taken abstractly to mean 'is at least as extreme as'. The p-value can be associated with a random variable P in its own right, defined by a function $f : X(\Omega) \to [0,1]$ (where $X(\Omega)$ is the image of the X) such that

$$P = f(X)|H_0 \sim U(0,1) \tag{1.3}$$

This second definition is useful when considering p-values from multiple tests, which may be considered as observations of P, and when reconstructing test statistics x as the (usually unique) pre-image $f^{-1}(p)$.

Under the Neyman-Pearson approach to hypothesis testing, the acceptance or rejection of H_0 is decided according to whether $p < \alpha$ for some predetermined α. An important quantity associated with this procedure is the type-1 error rate:

$$Pr(P < \alpha|H_0) \tag{1.4}$$

which, given equation (1.3), is equal to α. A second important quantity is the type-2 error rate, associated with a specific alternative hypothesis H_1 under which P has some known distribution $\neq U(0,1)$.

$$Pr(P > \alpha|H_1) \tag{1.5}$$

The complement of the type-2 error rate ($1 - Pr(P > \alpha|H_1)$) is also called the 'power' of the test, and is typically associated with a 'true' value of the original test statistic X. In most statistical applications, a testing procedure is chosen to control the type-1 error rate at α

while minimising the type-2 error rate. A type-1 error is the corresponding event in which H_0 is incorrectly rejected, and a type-2 error the event in which H_1 is incorrectly rejected.

The null-hypothesis significance testing procedure can be thought of as as a probabilistic application of Descartes' *Reductio ad absurdum*, which may have been what Laplace had in mind. To prove a statement in this framework, the contrary is assumed (H_0) and the researcher attempts to derive a contradiction, which is necessarily a probabilistic one when using observations from finite samples. The concept of quantifying evidence against a null hypothesis exemplifies the importance of falsifiability in the scientific method [Popper, 1957].

This framework of hypothesis testing has however come under criticism [Wasserstein and Lazar, 2016], generally relating to the difficulty of interpretation of the p-value. The p-value does not directly quantify the effect size of an association, although this may be reconstructible using both the p-value and the sample size. The p-value is also notably not equal to the probability that H_0 is true given observation x; indeed, this quantity has no meaning in the frequentist interpretation of probability.

The Bayesian interpretation of probability does allow meaning to be assigned to such events; namely the posterior probability of H_0 given X

$$Pr(H_0|X > x) = \frac{Pr(X > x|H_0)Pr(H_0)}{Pr(X > x)} \tag{1.6}$$

requiring the assignment of a prior $Pr(H_0)$ on H_0 before x is observed. The degree of support for a specific alternative H_1 over the null H_0 after observing X can be compared using a Bayes factor or likelihood ratio:

$$BF = \frac{Pr(X|H_1)}{Pr(X|H_0)} \tag{1.7}$$

The frequentist and Bayesian approaches to assessment of hypotheses each have advantages [Efron and Tibshirani, 2002] and are both readily susceptible to misinterpretation, but ultimately seek to answer different questions and are not generally comparable.

In this thesis, I use both frequentist and Bayesian interpretations of probability, generally due to the use of a combination of statistical methods. In chapter 2, a Bayesian probability is computed for each of a large number of tests, but the probabilities are then interpreted as test statistics in a frequentist multiple-testing procedure. It should generally be clear from the context which interpretation of probability is intended.

1.2.4 Increasing dimensionality

New scientific approaches

A common trend in many areas of applied statistics is a move toward higher-dimensional studies, with many variables being measured in parallel across a sample set, to the extent that the number of variables may substantially exceed the number of samples. A major driver of this trend has been a move towards 'omics' approaches [Bühlmann et al., 2014], in which all variables potentially contributing to the variance of some trait are exhaustively scanned for associations with the expectation that only a minority will be causally associated.

The omics approach has the advantage of near-agnosticism to prior hypotheses on association. This can be particularly useful in the analysis of complex traits in biomedicine, in which many variables may contribute small effects to the overall variance [Pearson and Manolio, 2008, Hirschhorn and Daly, 2005]. Targeted studies on a limited number of variables can be thought of as incorporating an implicit prior giving a vanishing weight to all variables not measured, which may be driven by incorrect assumptions on the biological pathways involved as well as being driven by cost and convenience considerations [Pearson and Manolio, 2008]. Implicit use may lead to high type-2 error rates.

A degree of subjectivity is still generally present in the analysis of genomic data. An important example relevant to the aims of this thesis is the assignment of SNP associations to local candidate causal genes, which is complicated by the possibility of SNPs modulating distal genes (trans-QTLs). An important area of current research is the formalisation of candidate gene selection; in chapter 3, an approach based on interaction DNA regions in relevant cell lines [Schofield et al., 2016] was used for this purpose.

An obvious difficulty of omics approaches is the management of type 1 error in the context of the large number of variables tested [Pearson and Manolio, 2008]. This means that a larger observed effect size is needed in order to differentiate an associated variable from a non-associated one. Importantly, this is not a disadvantage of omics approaches, but a consequence of preferring a non-informative prior across a greater number of variables.

In genomics, this difficulty is exacerbated in traits for which genetic causality is mainly determined by rare genetic variants. Variance in observed odds ratios is greater for rare variants, and consequently a larger observed odds ratio is needed to reject the null for a rare variant compared to a common one. Determining associations of rare variants is also made more difficult by the multiplicity of such variants in the genome, and greater susceptibility to quality-control problems [Hirschhorn and Daly, 2005]. The importance of rare variation in complex disease is an active topic of current debate [Morris et al., 2012].

Controlling type 1 errors

High-dimensional statistics requires careful control of errors, and determination of how error rates should be quantified. The choice of the best error rate to control may reflect the relative cost of type 1 and type 2 errors and the expected prevalence of true associations.

An analogue of the type-1 error rate if more than one hypothesis is tested is the family-wise error rate (FWER). If a series of independent tests against n null hypothesis H_0^1, H_0^2, ... , H_0^n are conducted, with resultant p-values represented by random variables P_1, P_2, ... P_n respectively (noting $P_i|H_0^i \sim U(0,1)$), the FWER is given by:

$$FWER = Pr(\bigvee_{i=1}^{n} P_i \leq \alpha_0 | H_0^1, H_0^2, ...) \tag{1.8}$$

To control the FWER at a value α, a p-value threshold α_0 on individual hypothesis tests may be attained by the Sidak correction:

$$\begin{aligned} \alpha_0 &= 1 - (1-\alpha)^{\frac{1}{n}} \\ &= \frac{\alpha}{n} + O(\alpha^2) \end{aligned} \tag{1.9}$$

where the first-order approximation represents the familiar Bonferroni correction.

However, the quantification of error rate using a single metric loses inherent meaning when more than one test is conducted, due to the absence of total ordering of Euclidean spaces of dimension > 1 compatible with the natural field structure[1]. The FWER, being the disjunction of all individual tests, is considered conservative as a metric; it effectively measures the probability of making at least one type-1 error.

An important drawback of the use of FWER-controlling methods is that they do not change with the observed distribution of p-values. For example, if a hundred parallel studies led to one of the two sets of p-values in figure 1.2, a Sidak correction would indicate a cutoff of $\approx \alpha/100$ to maintain a FWER $\leq \alpha$. However, the non-uniform distribution of p-values in the right panel would suggest that there are more non-null observations amongst the hundred than in the scenario on the left, and (in practical terms) we may attain more information from the data with a less stringent threshold for association.

[1]The absence of a total ordering on the plane in the consideration of bivariate test statistics will be reconsidered in chapter 1, when analysing the false-discovery rate of the cFDR method

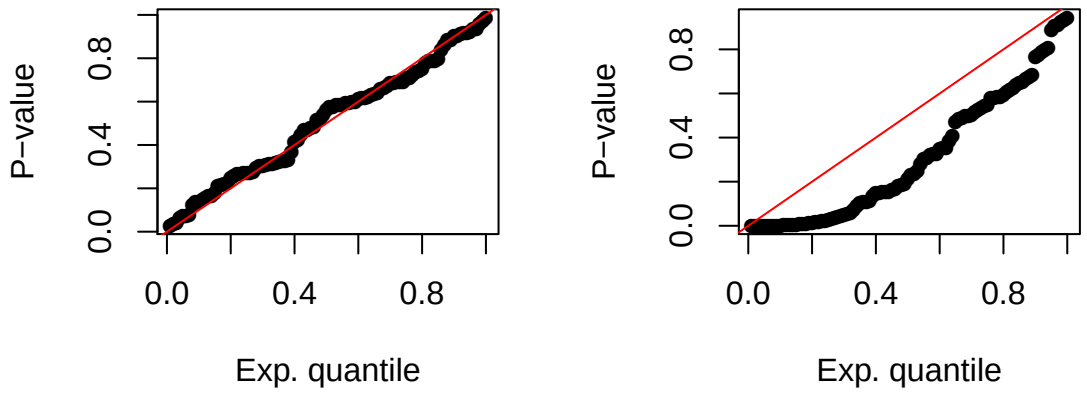


Fig. 1.2 Q-Q plots for two sets of observed p-values for which the associated random variable (equation (1.3)) has different distributions. Approaches to multiple hypothesis testing may or may not be responsive to such differences in distribution.

False discovery rates

The FWER considers only the probability of making 'at least one' type-1 error. Alternatives to the FWER generally involve considering the probabilities of making multiple type 1 errors, and limiting the rate at which errors are made. A useful such metric is the false-discovery rate (FDR) [Benjamini and Hochberg, 1995]. If a set of hypotheses are tested, R are rejected, and F are rejected falsely (that is, H_0 holds) the false-discovery-rate is defined as

$$FDR = E\left(\frac{F}{R}|R>0\right)Pr(R>0) \tag{1.10}$$

A useful theorem due to Benjamini and Hochberg [Benjamini and Hochberg, 1995] enables control of the FDR in a way which is responsive to the overall p-value distribution. If n independent or positively-dependent p-values p_1, p_2,... p_n are observed, and the relevant null rejected whenever

$$\frac{p_i}{\frac{1}{n}|\{j : p_j \leq p_i\}|} \leq \alpha \tag{1.11}$$

then the FDR associated with the procedure is bounded above by α.

A related quantity is the 'positive' false discovery rate (pFDR) [Storey, 2002] defined as

$$pFDR = E\left(\frac{F}{R}\Big|R>0\right) \tag{1.12}$$

Defining π_0 as the true proportion of null hypotheses that are false, definition 1.12 has the disadvantage that it is uniformly 1 if $\pi_0 = 1$, which can be problematic [Benjamini and Hochberg, 1995]. In this thesis, I will generally assume $Pr(R>0) \approx 1$, and use the FDR and pFDR interchangeably.

Bayesian interpretation

The quantity 1.12 may be interpreted in a Bayesian sense [Storey, 2002]. Assume m hypotheses $H_1, H_2..H_m \in \{0,1\}$ are tested by means of p-values $P_1, P_2...P_m$ and consider the H_i as iid Bernoulli-distributed random variables with $P(H=0) = \pi_0$. For a given threshold α, denote $R(\alpha)$ as the number of P_i with $P_i < \alpha$ and $V(\alpha)$ as the number of $P_i < \alpha$ with $H_i = 0$. If $P_i|H_i$ has CDF $F = F_0 H_i + F_1(1-H_i)$ and $F_1(x)/x$ is decreasing in x, then for fixed α we have [Storey, 2002]

$$\begin{aligned} FDR(\alpha) &= E\left(\frac{V(\alpha)}{R(\alpha)}\Big|R(\alpha)>0\right) \\ &= \frac{E\{V(\alpha)\}}{E\{R(\alpha)\}} \\ &= Pr(H=0|P<\alpha) \\ &= \pi_0 \frac{F_0(\alpha)}{F(\alpha)} \end{aligned} \tag{1.13}$$

The left-hand side of inequality 1.11 is an estimator of this quantity with $F_0(p_i) = p_i$, $\pi_0 \approx 1$, and $F(p_i) \approx \frac{\left|\{j:p_j \le p_i\}\right|}{n}$ [Efron et al., 2008]. This suggests the use of quantities of this form for each SNP as a marker of association in place of the p-value (termed the q-value) [Storey, 2002] an estimator for $Pr(H_0|P \le p_i)$,

A more useful posterior (the 'local' FDR [Efron et al., 2008]) evaluates the posterior probability of association for variables with p-value exactly α; defining f, f_0, f_1 as the PDFs corresponding to the CDFs F_0, F_1, F, the local FDR is defined as:

$$Pr(H_0|P=\alpha) = \pi_0 \frac{f_0(\alpha)}{f(\alpha)} \tag{1.14}$$

however, the estimation of f_0, f can be difficult. Potential methods include Lehmann alternatives (in which the assumption is made that $F_1(z) = F_0(z)^\gamma$) [Efron et al., 2008] or estimation of f from the dataset using empirical Bayes methods [Zablocki et al., 2014].

1.3 Genome-wide association studies

1.3.1 Motivation and development

The methods in this thesis were predominantly developed for use in datasets from genome-wide association studies (GWAS), and this is the application generally discussed.

Genome-wide association studies are attempts to discover the genetic causes of human traits in a hypothesis-free manner. They represent a generalisation of earlier candidate-gene studies (in which the effects of variants in putatively disease-associated genes are analysed in isolation), effectively involving computing multiple such studies in parallel and explicitly acknowledging the multiplicity of hypotheses being tested.

GWAS can contribute to medical knowledge and practice in several ways. An important motivation is the discovery of new disease-associated processes, and thus new targets for therapeutic drugs. A recent example is the identification of autophagy processes as a potential target in Crohn's disease [Zhang et al., 2008]

A second important application is the development of stratified medicine. Critical to this is understanding the causes and pathways driving disease heterogeneity, with the eventual aim of stratifying patients into sub-categories of disease. This may enable improved treatment efficiency by identifying patients in which treatments are more likely to be effective; for example, the identification of determinants of effectiveness of anti-TNFα therapy in multiple sclerosis [Gregory et al., 2012].

GWAS-driven patient stratification may also aid in identification of patients at risk of developing a disease; for example, the use of MHC-typing to identify patients at high risk of type 1 diabetes (T1D) [Noble and Valdes, 2011, Atkinson and Eisenbarth, 2001, Barrett et al., 2009]. Individuals with the highest-risk MHC genotype have a risk of developing T1D of approximately 5%, compared to a background risk of 0.3% in the general population [Aly et al., 2005]. Children identified as being at a high T1D risk can theoretically be more closely monitored for the onset of early signs and symptoms of the disease.

An extremal application of disease stratification is the development of personalised medicine; effectively the determination of the most effective treatment protocols on an individual scale. Although GWAS may not enable strong predictive ability, even when

disease heritability is high [Clayton, 2009], there may be scope for the modulation of individual treatment based on genetic risk assessment, although the field is controversial in its current form [Annas and Elias, 2014].

GWAS can be considered complete when the confirmed associations account for all of the observed the observed heritability (see section 1.3.4). For many common diseases, confirmed disease-associated variants only explain a small fraction of the overall genetic variance. There are many possible reasons for this including a high burden of rare variants [Fuchsberger et al., 2016], or a large number of common variants with small (non-zero) undetectable effect sizes [Golan et al., 2014], or a high burden of epistatic (non-additive) variation [Manolio et al., 2009]. Explaining all of the heritability of a trait may involve intractably large sample size; even for GWAS with sample sizes exceeding half a million, observed associations may not account for all of the variance of a trait [Wood et al., 2014]. A range of statistical methods have been developed for the analysis of rare variation [Lee et al., 2014], but in this thesis I generally only consider common variation (MAF $> 1\%$ or $> 5\%$).

As well as aiming to increase sample size, an important future direction of the GWAS field is an expansion into rarer and more specific traits. More specific phenotypes may have easier-to-observe associations, beneficial both for understanding such phenotypes and treating them.

In this chapter and throughout this thesis, I will refer to the 'genetic architecture' of a disease as the set of causal variants for that disease along with their effect sizes, in a typical British population unless otherwise stated. Occasionally I will use the term to instead mean the set of associated variants and their associated effect sizes; the specific meaning should be clear from the context.

1.3.2 Statistical methodology

In the GWAS design, a set of samples are genotyped at some large set of known variant sites (typically single-nucleotide polymorphisms, or SNPs) across the genome. In general, SNPs are assumed to be biallelic, meaning that the nucleotide at a given position in a chromosome may take one of two forms (of (A,C,G,T)). This thesis will generally consider autosomal (diploid) SNPs, for which the genotype of an individual may take one of three values: for alleles a, A, these are aa, AA (homozygous) and aA (heterozygous). For a variant under minimal selection and *de novo* mutation and with random mating, the frequencies $f(aa)$, $f(Aa)$, $f(AA)$ of aa, aA and AA in a population of size N satisfy

$$f(aa) = p^2 \qquad f(aA) = 2pq \qquad f(AA) = q^2 \tag{1.15}$$

as $N \to \infty$, where p is the frequency of allele a and $q = 1 - p$ the frequency of A (figure 1.3).

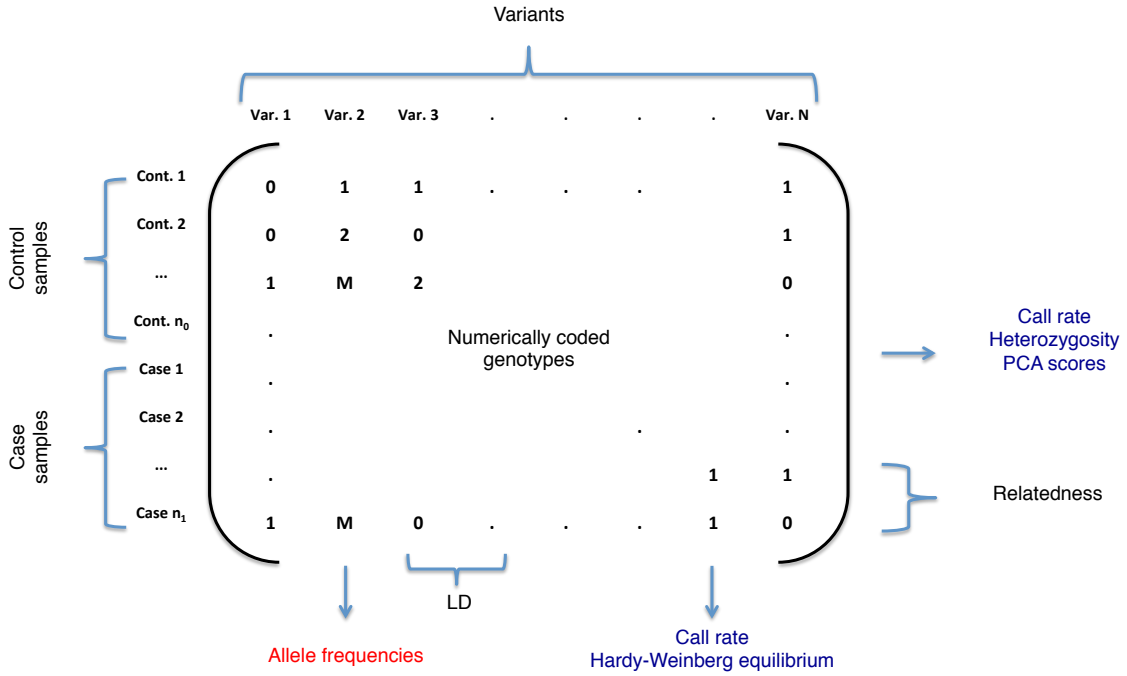


Fig. 1.3 At a SNP of interest, genotypes at biallelic SNPs are coded as 0,1,2, or M (missing) with heterozygosity corresponding to 1 and homozygosity in one allele (usually the rarer of the two) to 2. Genotypes at variants in close chromosomal proximity are often correlated, a phenomenon termed 'linkage disequilibrium'. Genotypes of samples may be correlated due to close relatedness. Column-wise averages (divided by two) from the submatrices for cases and controls are termed the allele frequencies, which serve as test statistics. Summary statistics derived from columns also include SNP-wise call rate (proportion of M's) and deviance from Hardy-Weinberg equilibrium, both used in quality control. Summary statistics derived from rows include sample-wise call rate, and heterozygosity rate. Principal components of the matrix (considering samples as observations of N random variables) are often used as covariates in genomic control (see section 1.3.3) and for identifying population structure

A case-control GWAS broadly compares observed allelic frequency between two groups of samples. Assume an observed allelic frequency m_0 in a 'control' group with $2n_0$ haplotypes and observed allelic frequency m_1 in a 'case' group with $2n_1$ haplotypes, with $E(m_0) = \mu_0$, $E(m_1) = \mu_1$ (see figure 1.3). In general, SNPs with two alleles are characterised by the lesser ('minor') of the two allele frequencies (MAF). Allelic difference is commonly characterised by either the direct difference $m_1 - m_0$ or the odds ratio

$$OR(m_1, m_0) = \frac{m_1(1-m_0)}{m_0(1-m_1)} \tag{1.16}$$

Several test statistics may be used to test the hypothesis $H_0 : \mu_0 = \mu_1$, the simplest being a contingency-table based test statistic (equivalent to a comparison of binomial proportions). If n_{ij} copies of allele j are observed in cohort i, the statistic is:

$$\sum_{\substack{i\in(case,control)\\ j\in alleles}} \frac{(n_{ij}-E(n_{ij}|H_0))^2}{E(n_{ij}|H_0)} = \left(\frac{m_1-m_0}{\sqrt{\frac{\bar{m}(1-m_0)}{2n_0}+\frac{\bar{m}(1-m_1)}{2n_1}}}\right)^2 \tag{1.17}$$

where $\bar{m}$ is the overall observed MAF, $\frac{n_0m_0+n_1m_1}{n_0+n_1}$, and the factor of 2 arises due to most chromosomes being diploid in humans (alternative methods are used for sex chromosomes). This test statistic converges in distribution to a χ^2_1 random variable, and hence an approximate Z score $\hat{Z}$ can be derived from the square root of this test statistic, with sign determined by the sign of $m_1 - m_0$. Under H_0, as $n_1, n_0 \to \infty$ with n_0/n_1 positive and finite, $\hat{Z}$ and each of the test statistics

$$\frac{m_1-m_0}{SE(m_1-m_0)} \qquad \frac{\log\{OR(m_1,m_0)\}}{SE(\log\{OR(m_1,m_0)\})} \qquad \frac{m_1-m_0}{\sqrt{\bar{m}(1-\bar{m})}\sqrt{\frac{1}{2n_0}+\frac{1}{2n_1}}} \tag{1.18}$$

(where *SE* denotes standard error) are asymptotically equal and converge in distribution to $N(0,1)$.

As with any case-control study, a major consideration is the potential for confounding variables also affecting μ_1, μ_0. For a vector z of covariates, a more appropriate null hypothesis may be $\mu_1(z) \equiv \mu_0(z)$, in which μ_0, μ_1 are considered as functions of covariates.

For categorical covariates, a Cochran-Mantel-Hanszel test is typically used - effectively a meta-analysis of a series of sub-analyses on each covariate level. For continuous covariates, tests typically use a logistic regression model on disease status Y:

$$\text{logit}(Pr(Y_i = 1)) = \alpha + \beta z + \gamma G_i + \varepsilon_i \tag{1.19}$$

where G_i is the genotype for sample i, β and γ are vectors corresponding to the covariate and genotype effects, and $\varepsilon_i \sim N(0, \sigma^2 I)$ is an error term. Another popular approach is a linear mixed model [Yu et al., 2006, Zhang et al., 2010, Lippert et al., 2011] in which $\varepsilon \sim N(0, KK'\sigma_g^2 + I\sigma_e^2)$, where K is a normalised version of the genotype matrix G of the form in figure 1.3:

$$K_{ij} = \frac{G_{ij}-2m_i}{\sqrt{2m_i(1-m_i)}} \tag{1.20}$$

where m_i is the observed allele frequency for SNP i. K is termed a 'genetic similarity' or 'kinship' matrix, and can also be estimated by other means [Speed et al., 2012],[Zhang et al., 2010].

The detection of variant-disease associations under the GWAS methodology generally requires sample sizes of several thousand. This large sample size is partly because of the dimensionality of the problem, and partly due to the nature of genetic architecture of common diseases. Variants with large effects on a disease phenotype are likely to be under negative selection, and hence at a low allelic frequency in the population. Large GWAS on common diseases [Barrett et al., 2009, Morris et al., 2012, Okada et al., 2014] have nonetheless found a large proportion of disease heritability to arise from common (high MAF) variants. The question of the relative contributions to genetic disease risk from low-frequency variants of large effect and from high-frequency variants of low effect is a continuing debate [Gibson, 2012] but both genetic architectures are problematic in that the types of variants responsible for disease causality are statistically difficult to find.

1.3.3 Quality control

A vital part of GWAS is ascertainment of data quality. Studies are susceptible to a multitude of potential confounders, both due to subtle effects on genotype accuracy due to sample storage or processing and differential population sampling, and failure of type-1 error rate control is costly due to unnecessarily wasted resources in downstream analyses. Many aspects of GWAS quality control are specific to the particular study at hand [Anderson et al., 2010], and hence development of new methodology such as that in this thesis must include consideration of appropriate quality control procedures. I include an introduction here to several standard procedures for conventional GWAS.

Quality control in GWAS is generally involves both systematic procedures across all variants, and specific procedures on variants identified as putatively associated. For analyses of patterns of effect sizes across large numbers of SNPs, only the first of these is possible, and hence must be somewhat more stringent. This is relevant to chapters 5 and 6, in which genome-wide effects are quantified.

Systematic quality control effectively has three main steps - the exclusion of variants, the exclusion of samples and the management of residual inflation. Variant exclusion is typically on the basis of poor call-rate, low allele frequency or deviation from Hardy-Weinberg equilibrium [Anderson et al., 2010]. Sample exclusion is typically on the basis of poor call

rate, incongruence of reported sex (a mismatch between observed and expected heterozygosity on chromosome X), or ethnicity outside of the study population (see figure 1.3).

Confounder-driven inflation in test statistics after removal of samples and variants in this way is usually quantified using the formula [Devlin et al., 2001]

$$\lambda = \frac{\text{median}(\chi_1^2(i))}{Q_{\chi_1^2}\left(\frac{1}{2}\right)} \tag{1.21}$$

where $\chi_1^2(i)$ is a χ^2 test statistic for SNP i and $Q_{\chi_1^2}$ is the quantile function of the χ^2 distribution with 1 degree of freedom . This may not be appropriate if a high proportion of the genome is associated with the trait [Wood et al., 2014] or in GWAS-like analyses on only areas of the genome with a high prior for association [Cortes and Brown, 2011, Stahl et al., 2010]. In this case, the inflation factor λ may be estimated by only using a set of variants specifically chosen so that H_0 is expected to hold [Eyre et al., 2012].

Management of inflation may involve accounting for more covariates, either by further stratifying samples or inclusion of more covariates in a logistic model. A common approach is to include several principal components of the genotype matrix as covariates (where principal components correspond to weighted sums of alleles) [Price et al., 2006]. Principal component covariates index the main axes of polygenic variation in the dataset, which are usually assumed not to be the sources of variation under investigation in the study; however, the use of principal components as covariates cannot generally correct for unexplained (cryptic) relatedness between individuals in a study. An alternative way to manage inflation from population structure and manage cryptic relatedness at the cost of some statistical power is to fit a linear mixed model [Zhang et al., 2010] (see section 1.3.2). Both approaches are imperfect: they may not completely control type 1 error (that is, not remove all inflation) and may induce type 2 error by 'blindly' correcting for true genetic signals, especially on very polygenic traits such as height. Residual inflation after these steps is typically removed by simply dividing observed χ^2 statistics by the inflation factor λ, with the effect of forcing the value of λ to 1. Since this procedure also reduces the observed effect size at true associations, it reduces the power of the study, so it is generally desirable to reduce λ as much as possible before this step.

1.3.4 Heritability and genetic correlation

It is possible in many cases to infer useful information from GWAS without the need to specifically identify disease-associated variants. One important metric in genomics is the

heritability, which is colloquially considered 'the proportion of variance in a trait attributable to genetic variation'.

In a more formal sense, for a trait Y with non-zero prevalence in a population P, let G denote a random variable associated with the complete genotype of an individual in P. The heritability H^2 is defined as

$$H^2 = \max_{\substack{f:G\to\mathbb{R} \\ \mathrm{var}(f(G))=1}} \lim_{|P|\to\infty} \mathrm{cor}\,(f(G),Y) \tag{1.22}$$

The narrow-sense heritability h^2 restricts f to linear functions of G [Bulik-Sullivan et al., 2015], and the SNP heritability further restricts G to some subset of the SNPs in the genome (typically those on a genotyping chip). Both wide and narrow-sense heritability depend on the population being sampled and the environment that population is in.

The definition of h^2 invites an interpretation of the genetic architecture of a disease as a point in a high-dimensional space, with the co-ordinate β_i of a variant i corresponding to the proportion of $\mathrm{var}(Y)$ attributable to it. The correspondence of the value

$$h^2 = \sum \beta_i^2 \leq 1 \tag{1.23}$$

to the 2-norm of the vector invites consideration an inner product on this space characterising the similarity between two traits; the obvious such being the 'genetic covariance', ρ_g between vectors α, β:

$$\rho_g = \sum \alpha_i \beta_i \tag{1.24}$$

usually considered in the form of the 'genetic correlation'

$$r_g = \frac{\sum \alpha_i \beta_i}{\sqrt{\sum \alpha_i^2}\sqrt{\sum \beta_i^2}} \tag{1.25}$$

The genetic correlation measures pleiotropy (association of the same variant with > 1 phenotypes) between α and β in a sense, but importantly responds differently when effects are in the same direction for both traits ($\alpha_i\beta_i > 0$) than when they are both nonzero but in opposite directions ($\alpha_i\beta_i < 0$). This distinction is discussed further in chapter 5 in the context of characterising shared and distinct genetic associations with disease subtypes.

Common methods for estimation of SNP h^2 and r_g include estimates from linear mixed models (implemented as 'genome-wide complex trait analysis' (GCTA) [Yang et al., 2011]) and LD-score regression (LDSC) [Bulik-Sullivan et al., 2015] . In the model implemented in

GCTA, the trait y is modelled as

$$y = Wu + X\beta + \varepsilon \tag{1.26}$$

where W is a matrix of numeric genotypes normalised so each row has zero mean and unit variance, u is a vector of random effects with variance $\sigma_u^2 I$, X is a matrix of covariates with fixed effects β, and $\varepsilon \sim N(0, \sigma_e^2 I)$. Given

$$\text{var}(y) = XX^t\sigma_u^2 + \sigma_e^2 I \tag{1.27}$$

the variance components σ_u, σ_e can be estimated by the restricted maximum likelihood approach [Patterson and Thompson, 1971] with the estimate of narrow-sense heritability given by $N\sigma_u^2$, for N SNPs.

If two traits y_1, y_2 are measured in some population of size n on N SNPs (possibly with missing observations, as for two separate studies), narrow-sense heritabilities [Yang et al., 2011] and genetic correlation [Lee et al., 2012] between y_1 and y_2 can be estimated using a bivariate LMM:

$$\begin{aligned} y_1 &= Wu_1 + X\beta_1 + \varepsilon_1 \\ y_2 &= Wu_2 + X\beta_2 + \varepsilon_2 \end{aligned} \tag{1.28}$$

where X is a matrix of covariates with fixed effects β_1, β_2 in y_1 and y_2, W is a $N \times n$ matrix of numeric genotypes normalised so each column has zero mean and unit variance, u_1, u_2 are vectors of random effects and ε_1, ε_2 vectors of random residuals with

$$\text{var}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} I_N\sigma_{u1}^2 & 1_N\sigma_{u12} \\ 1_N\sigma_{u12} & I_N\sigma_{u2}^2 \end{pmatrix}, \quad \text{var}\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = \begin{pmatrix} I_n\sigma_{e1}^2 & 1_n\sigma_{e12} \\ 1_n\sigma_{e12} & I_n\sigma_{e2}^2 \end{pmatrix}$$

$$\text{var}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} WW^t\sigma_{u1}^2 + I_n\sigma_{e1}^2 & WW^t\sigma_{u12} + I_n\sigma_{e12} \\ WW^t\sigma_{u12} + I_n\sigma_{e12} & WW^t\sigma_{u2}^2 + I_n\sigma_{e2}^2 \end{pmatrix} \tag{1.29}$$

The variance components σ_{u1}^2, σ_{u2}^2, σ_{12}, σ_{e1}^2, σ_{e2}^2, σ_{e12} can be estimated by the restricted maximum likelihood approach [Patterson and Thompson, 1971]. Estimates of narrow-sense heritability for y_1 and y_2 are given by $N\sigma_{u1}^2$, $N\sigma_{u2}^2$, and an estimate of genetic correlation is given by $\sigma_{u12}/(\sigma_{u1}\sigma_{u2})$. The model specification $u \sim N(0, \sigma_u^2)$ implies that all SNPs contribute to the trait, which is inconsistent with several models of genetic architecture [Gibson, 2012]. Alternative models inducing sparsity have also been proposed [Zhou et al., 2013].

GCTA can underestimate narrow-sense heritability when applied to case-control studies in which the prevalence of cases in the study is higher than the prevalence of cases in the general population [Golan et al., 2014].

The LDSC method is an alternative approach which uses the assumption that causal variants in LD with each other will have inflated observed effect sizes [Bulik-Sullivan et al., 2015]. If the LD between SNPs i and j is r_{ij}^2, the LD-score for a SNP i is defined as

$$l_i = \sum_{\text{SNPs } j} r_{ij}^2 \tag{1.30}$$

Define z_{1i}, z_{2i} as the z-scores for SNP i for quantitative traits y_1 and y_2 on N_1, N_2 samples respectively, with N_s samples in common and M SNPs in total. The narrow-sense genetic correlation ρ_{12} between y_1, y_2 satisfies

$$E(z_{1i}z_{2i}) = \frac{\sqrt{N_1N_2}\rho_{12}}{M} + \frac{N_s}{\sqrt{N_1N_2}}\text{cor}(y_1, y_2) \tag{1.31}$$

A similar equation holds for case-control studies. If the two studies are the same, then $N_1 = N_2 = N_s$ and ρ_{12} is the heritability h^2 of the trait, and hence

$$E(z_i^2) = \frac{N}{M}h^2 + \text{var}(y_i) \tag{1.32}$$

By regressing LD-scores on z-scores, the heritability and narrow-sense genetic correlation can thus be estimated. LDSC typically needs more samples than GCTA, but only requires summary statistics for the estimation procedure [Bulik-Sullivan et al., 2015].

Broad-sense heritability (H^2) is typically estimated from twin studies, which are outside the scope of this thesis.

1.3.5 Genetic risk scores

As well as determining the set of variants associated with a trait of interest, it can be useful to attempt to recover the underlying (usually additive) genetic risk model. Models derived in this way can give rise to 'genetic risk scores' (GRS) which function as a invariant lifetime risk predictor of the trait.

Unfortunately, predictive accuracy of such a score may be poor even when heritability is high [Clayton, 2009]. However, in some cases GRS can contribute meaningfully to clinical disease prediction, making use of the polygenicity of common traits [Abraham et al., 2014].

Under the additive risk model of narrow-sense heritability, a genetic risk score for individual i with genotype $G_s(i)$ at SNP s typically takes the form

$$S(i) = \sum_{s \in genome} \beta_s G_s(i) \tag{1.33}$$

where β_s is the coefficient of SNP s in the model, typically an estimator for the true effect size of the SNP. The values β_s may be 0 for most s if a sparse model is needed.

Genetic risk scores may be fit only using known disease-associated SNPs [Yarwood et al., 2015] or on SNP sets for which only a proportion are assumed to be disease-associated [Abraham et al., 2014].

In the latter case, machine-learning methods are typically used to impose sparsity and prevent overfitting. A popular choice is lasso (L1-penalised) regression (for example [Barrett et al., 2009]), although non-linear approaches such as support-vector machines may be more effective in some circumstances [Abraham et al., 2014].

1.4 Comparison of association studies

1.4.1 General considerations

The history of association studies has been marked by the assembly of similar cases, and analysis of similarities of the resultant cohorts. As discussed earlier, a major disadvantage of this approach is the necessary agnosticism to phenotypic heterogeneity in order to attain large enough sample sizes. This shortcoming has been recognised since the first such studies [Feldman and Goodrich, 1999].

This invites the possibility of learning more from GWAS by subdividing diseases. In the opposite direction, there may be information to be gained by comparing studies on related diseases together. This is made practically difficult by variant definitions of phenotypes and a wide range of genotyping platforms. Another major obstacle is reluctance of research groups to share data [Piwowar et al., 2008] both for reasons of patient anonymity and monopoly on research outputs. For the former reason, frequently only summary statistics are made available from GWAS rather than genotypes (eg [ImmunoBase, 2013]), and available summary statistics may not include effect directions.

Both of these shortcomings in disease heterogeneity in GWAS studies and the difficulties with comparisons of association analyses are gradually being alleviated; the former by enlargement of sample cohorts and improved phenotypic characterisation (eg [UK Biobank,

2007]), the latter by homogenisation of association study design and independent variables (especially in genomics) and better-facilitated data-sharing. This has necessitated a range of statistical methods for comparing association studies.

Such methods may seek to answer several questions. Denoting H_0^A/H_1^A, H_0^B/H_1^B, X^A, X^B as null/alternative hypotheses and test statistics for association with diseases A, B respectively, a commonly estimated (Bayesian) probability is

$$Pr(H_1^A \vee H_1^B | X^A, X^B) \tag{1.34}$$

Computation of this quantity effectively corresponds to a meta-analysis of A and B as if they were the same disease (so $H_0^A = H_0^B$) and hence it is of limited use if $A \neq B$ and individual association with A and B is of interest rather than association with $A \vee B$. A more useful quantity in many cases is

$$Pr(H_1^A \wedge H_1^B | X^A, X^B) \tag{1.35}$$

with the aim to identify shared association. This quantity is of use in the two-stage replication procedure [Wason and Dudbridge, 2012] when A and B nominally represent the same phenotype; this is discussed in chapter 4. Finally, the asymmetric posteriors

$$Pr(H_1^A | X^A, X^B) \qquad Pr(H_1^B | X^A, X^B) \tag{1.36}$$

are useful in 'leverage'; using information about one disease to inform another. This is discussed in chapters 3 and 6. Assessing the pattern of distribution of associations between two phenotypes is a more complex problem, one aspect of which is discussed in chapter 5

The optimal procedure for comparing association studies depends on the aims of the comparison. In particular, procedures differ according to whether the two studies being compared are on the same nominal disease, in which case the sets of associated variables in each study are assumed to be the same, or on different diseases, in which case some associations may be unique to one of the studies being compared. This section is organised on this basis.

Comparison of association studies is frequently done in an informal way; for instance, observing that the same region appears to be associated with several different diseases [The Wellcome Trust Case Control Consortium, 2007]. For low-dimensional association studies with reasonable power, this has sufficed; indeed, as discussed in section 1.2.2, intuition historically sufficed for all analyses of disease cohorts until relatively recently. However, in the same way that modern medicine has necessitated more complex studies and more precise

analysis, the future of medicine may come to require equivalent standards of precision in the co-analysis and comparison of studies.

1.4.2 Studies of different diseases

Identifying pleiotropy

An important question in joint analysis of two disease phenotypes is determination of which variables are associated with both of them. This is of interest in the understanding of disease pathophysiology, as a shared association indicates some degree of sharing in the causal pathways of both diseases. It is also useful in designing therapeutic approaches, as a shared association suggests drugs targeting a shared pathway may be of use in the treatment of both phenotypes; for example, the recent investigations in the use of anti-inflammatory drugs in depression [Köhler et al., 2014].

In the context of GWAS, the phenomenon of shared associations is broadly termed 'pleiotropy'. Throughout this thesis, I will consider a variant to be pleiotropic for two diseases if it is associated with both (without necessarily being causal for both) and two diseases to 'show pleiotropy' if their corresponding sets of associated variants overlap. This definition is common in GWAS and related studies [Andreassen et al., 2013] although the term does not have a fixed meaning and various interpretations are used in the field according to context [Paaby and Rockman, 2013]. The definition above necessitates an acceptance of the 'two-groups' model [Efron et al., 2008] in which variants may be partitioned into those associated with the disease and those not associated, usually with the implicit assumption that most variants are in the latter category. The definition used in this thesis is troublesome under the 'universal pleiotropy hypothesis' [Paaby and Rockman, 2013] in which all variants have a small effect on all phenotypes; for example, the random-effects model in [Yang et al., 2011].

Pleiotropy is common in human diseases: even when exclusively considering single nucleotide polymorphisms (SNPs) with strong evidence of association, around 15% of those associated with at least one trait are associated with multiple traits [Sivakumaran et al., 2011]. Elements of shared genetic aetiology may be suspected in diseases with similar symptomatology, such as bipolar disorder and schizophrenia [Lichtenstein et al., 2009] or in diseases with common risk factors, such as type 2 diabetes and obesity [Hasstedt et al., 2011]. If two diseases are known or suspected to share associated genetic variants, a degree of association of a locus with one disease may increase the likelihood of association with the other. This suggests that incorporating datasets from diseases related to the disease of

interest may help alleviate some of the effect of multiple testing, meaning that phenotypic similarity may lead to improved detection of disease-associated variants [Galesloot et al., 2014]. Correspondingly, discovery and specification of shared genetic aetiology between two diseases may suggest some shared pathophysiology [Hasstedt et al., 2011].

The genetic correlation of two traits (equation (1.25)) quantifies pleiotropy in a sense, but can be misleading if effect sign differences are inconsistent between variants (discussed in chapter 5). Non-zero genetic correlation does, however, provide sufficient evidence for pleiotropy between two traits, and there is evidence of widespread pleiotropy between many common autoimmune and metabolic diseases using this method [Bulik-Sullivan et al., 2015].

On a SNP-by-SNP basis, there are several approaches to the determination of shared association with more than one phenotype [Galesloot et al., 2014, Shriner, 2012], a common theme to many being the combination of multiple traits into a single generic trait to reduce dimensionality [O'Reilly et al., 2012, Klei et al., 2008]. Most methods are designed for the case when all phenotypes are measured on the same individuals, although some can be be adapted to circumstances where this is not the case. However, when analysing multiple studies on different individuals, the most widely applied approach is to simply compare the sets of variants reaching genome-wide significance (GWS) in two studies (for instance, [The Wellcome Trust Case Control Consortium, 2007]), which suffices for large effects although is problematic if pleiotropic variants have markedly different effect sizes between diseases. A similar method is to specifically analyse SNPs associated with one disease of interest in a GWAS for a separate disease, typically using a somewhat relaxed threshold for significance [Plagnol et al., 2011, Ramos et al., 2011]. Both of these methods are susceptible to inflated type-1 error rates if samples are shared between studies and appropriate corrections are not made [Bhattacharjee et al., 2012].

A similarly important question in the symmetric assessment of pleiotropy is whether two diseases share a *causal* SNP in a region of interest (termed 'colocalisation'), a stronger statement than the sharing of association. The assessment of colocalisation is a complex procedure which may be approached in a Bayesian model averaging framework [Giambartolomei et al., 2014, Fortune et al., 2015]. Methods are not discussed in this thesis; however, an assertion of pleiotropy in a region must be further assessed with colocalisation analysis before conclusions on causality can be drawn.

Leverage

Even if we are unable to assert association with both of two phenotypes of interest, joint analysis of both diseases may be useful in identifying association with each individual

disease. As an example, assume that for two phenotypically related phenotypes A and B, independent GWAS found p-values $P_A = 1 \times 10^{-7}$, $P_B = 1 \times 10^{-4}$ against H_0 at some variant. A meta-analysis of A and B (as though the diseases were identical) would probably conclude that the variant was likely to be associated with $A \vee B$; but this does not entail association with $A \wedge B$. While neither p-value is small enough to assert GWS at the standard threshold ($p < 5 \times 10^{-8}$), we may intuitively assert that there is good evidence for association with A, given the additional low p-value for B. However, the intuitive evidence for association with B is considerably poorer.

In this way, more information may be extractable from a joint dataset by considering association with each disease individually, potentially gaining some of the advantages of a meta-analysis in power while avoiding compromising specificity of the phenotype, a process which I generally refer to in this thesis as 'leverage'. The most widely-used method of leverage is probably the analogy of the method described in the section above to demonstrate pleiotropy: the restriction of variants analysed for disease B to those with some level of association with disease A. However, this approach loses information from variants associated with disease A at sub-significant effects, which may still contribute useful information [Park et al., 2010] (as in the example above). A useful extension to this method is the Bayesian conditional false discovery rate (cFDR) which uses an analogy of the Benjamini/Hochberg procedure to two phenotypes [Andreassen et al., 2013]. This is discussed in detail in chapter 2.

An important application of association study comparison is in the analysis of rare diseases. The field of GWAS and other omic- studies has the advantage of being hypothesis-free at the cost of power. For rare phenotypes, the assimilation of the thousands of cases typically necessary to power a GWAS is often impractical, and GWAS are of limited use. In order to analyse the phenotypes at all, some leverage on larger case datasets needs to be performed. Chapters 3 and 6 describe several procedures of this type.

1.4.3 Studies of the same disease

Meta-analysis

The procedure of meta-analysis is important in the systematic assessment of evidence. Meta-analytic procedures have been developed for genetic association studies, and has enabled effective sample sizes of $> 10^5$ for some traits [Wood et al., 2014]. In a contrary trend to that of finer phenotype specification, it is also frequently useful to meta-analyse different but related phenotypes, estimating the quantity in equation (1.34), generally followed by analysis of the individual associations with the constituent phenotypes [Li et al., 2015].

An important consideration in meta-analysis of association studies is the presence of shared control samples between groups. Several methods for adapting meta-analyses to shared-control designs have been developed [Lin and Sullivan, 2009, Bhattacharjee et al., 2012], which provided a basis for the methods developed in chapters 2 and 4.

GWAS replication

A common exercise in high-dimensional association studies is replication, which has became standard practice in GWAS [Anderson et al., 2010]. The procedure typically involves conducting an initial study A on a disease of interest, and re-analysing putative associations in a second follow-up study A'. Given P-values P^A and $P^{A'}$ for A and A' and P for a meta-analysis, a typical replication procedure [Wason and Dudbridge, 2012] would declare a variant associated with A if

$$P^A \leq \alpha \qquad P^{A'} \leq \beta \qquad P < \gamma \tag{1.37}$$

for some thresholds α, β, γ, where γ is typically the GWS threshold, 5×10^{-8}. Effect directions are also required to be consistent across A, A'.

The replication procedure has the advantage of increasing the type-1/type-2 error rate profile to that of a meta-analysis for variables measured in both cohorts, and is cheaper since not all variables are generally measured in A'. Given the extensive potential for confounders and other errors in GWAS, the requirement for evidence of association in both A and A' is generally considered necessary [Anderson et al., 2010], with the understanding that unrecognised confounders affecting A will not affect A', and vice versa.

This is easiest to ensure if A and A' are independent; that is, share no samples. However, in some cases, it may be advantageous to share controls (or share one group) between studies. This is explored in chapter 4.

1.4.4 Investigating heterogeneity

As discussed previously, GWAS analysis of complex phenotypes is difficult due to a typical genetic architecture of rare variants with large effects or common variants with small effects [Gibson, 2012]. This difficulty is prominent in syndromes with diagnosis based on symptom clusters, such as schizophrenia [Ripke et al., 2014]. Accounting for phenotypic heterogeneity can improve the power of association analyses in these circumstances [Morris et al., 2009].

However, a division of a disease into subtypes will only yield information over an analysis in which the disease is considered homogeneous if effect sizes for disease-associated variants are different between the subtypes. More generally, the best mode of analysis for a heterogeneous disease will depend on the relations between the genetic architectures of the disease subtypes, which in turn relates to the underlying cause of disease heterogeneity. The joint genetic architectures of two disease subtypes can be modelled as a bivariate distribution of effect sizes, which gives a framework in which different causes of disease heterogeneity can be distinguished. This is discussed in chapter 5. In chapter 6, I analyse the genetic architecture of a disease with more than two subgroups using genetic risk scores of various types.

Searching for subtypes

The process of identifying genetic differences amongst subtypes of a disease invites the question of whether the process may be inverted to find clinically interesting subtypes of a disease in a genetically-driven way. There has been success in this area using multivariate clinical data [Siroux et al., 2014] and transcriptomic data [Bullinger et al., 2004] but it is a difficult problem using genomic data.

Attempts to find 'maximally separated' subtypes in a phenotype is difficult due to the diversity of causes of variation in the genome. Many observable variant human traits have some degree of heritability; including quantifications of anthropometry, behaviour and metabolic state [SNPedia, 2017]. Finding a 'maximally separated' partitioning of a case group thus broadly corresponds to finding the trait for which the genetically-driven variance is greatest, and this is very unlikely to correspond to the variance in disease pathology; it should generally be population structure (non-uniform patterns of identity-by-descent). This is supported by the regular use of principal components as quality-controlling covariates in GWAS logistic regression models [Price et al., 2006]; PCs characterise the axes of maximum variance within a dataset, and generally correlate strongly with ethnicity [Anderson et al., 2010, Price et al., 2006].

This problem may be alleviated by either restricting the set of variants considered to those likely to correspond to variation in the disease of interest, or restricting the space of potentially allowable subdivisions of a sample cohort into subtypes, which would otherwise have size 2^N for N samples. An obvious way to find such variants is to consider those with confirmed or putative association with the disease, but clinically-important subtypes may be differentiated by genetic variants not associated with the disease under investigation [Lee et al., 2017]. Other options are to use variants associated with a second disease, an approach

suggested in [Han et al., 2016b]. Chapter 6 describes an investigation into 'meaningful' subtypes of juvenile idiopathic arthritis (JIA) using both of these restrictions.

1.4.5 Efficient use of data

The general theme of this thesis is an aim to increase the amount of information attained from a given dataset, allowing greater economy of sample size. Historically, this has been relevant for all GWAS studies, due to low sample sizes and restricted access to data (for instance, GWAS summary statistics only). As datasets have grown larger, the use of efficiency - gaining methods such as sharing control samples or making use of leverage has become relatively less useful.

In the current setting and in the future, however, the trend toward finer and finer phenotypes may mean that studies on small cohorts remain common. It is frequently useful to compare small subgroups of samples against each other (for example, [Lee et al., 2017, Plagnol et al., 2011]; also see chapter 5), and in such studies shortcomings in power will arise both from undersized 'case' and 'control' cohorts. In this setting, I believe that statistical options to improve efficiency of data use and reduce the cost of recruitment are likely to remain important.

Shared controls

Particular care is required in comparison of association studies if the studies being compared have samples in common; for example, controls sourced from a common population source [The Wellcome Trust Case Control Consortium, 2007]. If controls are shared between studies on two traits, observed effect sizes between the studies will be correlated even for variants not associated with either trait [Bhattacharjee et al., 2012]. A common solution is to simply split the available control cohort into two independent sub-cohorts, each of which is only used for one of the two studies [Andreassen et al., 2013, Yang et al., 2011]. This procedure is is only possible if individual genotypes are available. More importantly, however, this 'split-control' approach is inefficient. In many study designs in which either shared controls or 'split controls' are usable [Skol et al., 2006, Bhattacharjee et al., 2012, Fortune et al., 2015, Lin and Sullivan, 2009], splitting a control group leads to a loss of power at the same type-1 error rate (also see chapters 2 and 4).

The inefficiency of splitting a control dataset is predictable from a heuristic argument. Assume that two association studies are performed, one comparing cohort C_1 to cohort C_0, the other comparing C_1' to C_0', where C_0 and C_0' are drawn from an identical control population.

Denote by m_1, m_0, m_1', m_0' the observed values of a variable of interest (for instance, allele frequencies) in each cohort. In a split control design, the association statistics depend on m_1, m_0, m_1', m_0' only through the differences

$$m_1 - m_0 \qquad\qquad m_1' - m_0' \tag{1.38}$$

This makes no use of the information that $E(m_0) = E(m_0')$ - whereas a shared control design, with C_0 and C_0' pooled together, does make use of this information.

In general, if association studies are to be compared and it is reasonable to assume (in the above notation) that $E(m_0) = E(m_1)$, procedures enabling controls to be shared should be considered. This generally requires some modification of statistical methods.

Information in a genotype matrix

A second important theme of this thesis is the comparison of motives for combining case groups ('lumping') and for separating case cohorts into different subtypes ('splitting'). This is a commonly recognised dichotomy in biomedicine [McKusick, 1969], possibly dating to Charles Darwin [Simpson et al., 1945]. In general, the best approach to take depends on the circumstances of the study, including the sample size, the phenotype in question, the study design, and the study objectives.

An important consideration is that GWAS and similar studies are highly reductionist, using only a small proportion of the available input information. Considering a standard genotype matrix G with each row corresponding to a sample, a GWAS effectively only compares allele frequencies, or column-sums. This only encompasses a linear-order proportion of the quadratic-order information the genotype matrix contains. Although information independent of these column-sums is used for quality control (call rates, PCA) and in pruning by LD, it is not generally used to directly modify the study design; for instance, by suggesting subgroup analyses.

There are many forms a genotype matrix may take while holding column-sums constant. For instance, if the phenotype consists of two completely non-pleiotropic subtypes, the genotype matrix will have a markedly different form to that of a cohort from a homogeneous phenotype, even though the column-sums (and hence GWAS results) may be the same (figure 1.4). The form of the matrix will also vary depending on the degree to which the phenotype is governed by common variants with small effect or rare variants with large effect [Gibson, 2012].

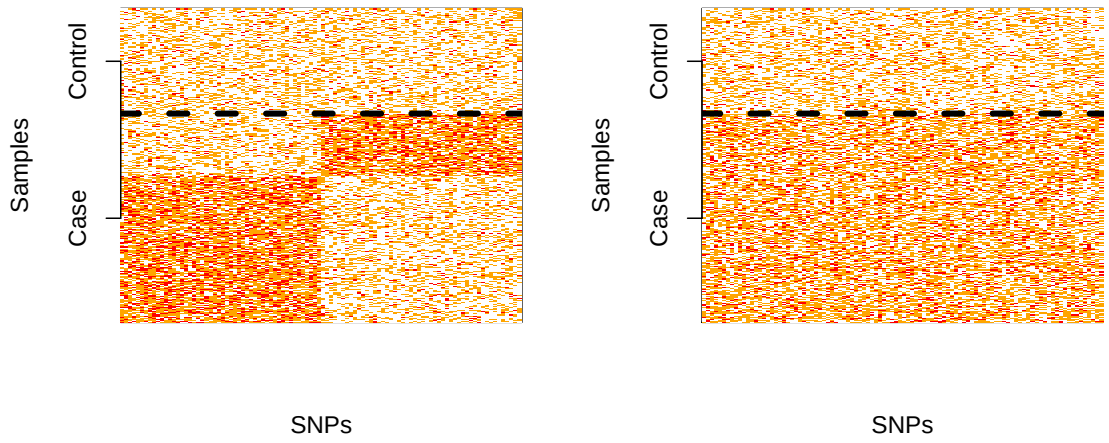


Fig. 1.4 GWAS analyses based only on allele frequencies use only a small proportion of the information the matrix contains. This figure demonstrates two genotype matrices with the same case/control allelic frequencies, and hence the same result from a GWAS analysis, but differing substructures. In the matrix on the left, the genetic architecture of the disease consists of two independent sets of associations, each set affecting one underlying subtype of the trait. A GWAS analysis taking the subtype structure into account would be more powerful than an analysis of the trait as a whole. However, in the matrix on the right, no such substructure is present (the phenotype is homogeneous), and analyses accounting for subgroups of the case group would only introduce noise, potentially lowering power.

The variation of GWAS methodology on the basis of the structure of the genotype matrix may enable the use of more information. This is difficult to do in a general sense as discussed in section 1.4.4, but when used appropriately could enable more powerful analyses than conventional GWAS on the same datasets, possibly with incorporation of external information. Examples of usage of external and column-sum independent information to understand disease structure are discussed in chapters 5 and 6.

1.5 Overview of chapters

As discussed in the foreword to this introduction, chapters 2-6 describe a series of projects with common themes rather than a continuous narrative. Two chapters (chapter 2 and chapter 5) have been published, and two others (chapter 4 and section 3.3 of chapter 3) have each undergone at least two rounds of peer-review. In all cases the chapters in question are not identical to the publications or submissions, but they do contain transcribed passages from papers for which I wrote the first draft. Sections from manuscripts may have been modified by co-authors from my original drafts in wording, but not in substance. I was the sole author for the manuscript based on chapter 3. In general, all work presented in this thesis and its appendices is my own unless directly specified.

Chapter 2 describes the extension of a Bayesian conditional false discovery rate to a shared-control design, and describes an approximate bound on overall false-discovery rate when using the procedure. The chapter also includes an application to ten autoimmune diseases, considered two at a time. The chapter was published in [Liley and Wallace, 2015].

Chapter 3 describes two further applications of the cFDR method: firstly, to (JIA), conditioning on T1D and RA, and secondly to a rare form of vasculitis, EGPA, conditioning on asthma and physiological eosinophil count. This chapter also serves as an exploration of the practicalities of applying the cFDR method. The second part of this chapter (on EGPA) is under review ([Lyons et al., 2017], on which I am joint first author).

Chapter 4 describes a different application of the shared-control design in the context of study replication. This includes suggested applications of the new method. The chapter underwent two rounds of peer-review at the journal Genetic Epidemiology, but was ultimately rejected (with potential to resubmit) on the grounds of the absence of a specific application. It is currently on arXiv [Liley, 2017].

Chapter 5 describes an exploration into the characterisation of genetic architectures of disease subtypes, with application to T1D. This chapter incorporates an application of the

cFDR method and is itself a potential application for the methods discussed in chapter 4. This chapter was published in [Liley et al., 2016].

Chapter 6 describes recent work into the subtypes of JIA. It is similar in aim to chapter 5, but differs in several ways. The analysis is largely exploratory but includes discussion of predictability of arthritic extension (worsening) and the relative genetic similarity of the seven major JIA subtypes.

Chapter 7 is a discussion of the linking themes in this thesis, and includes some suggestions for future work. Each chapter includes a more specific discussion with a more technical focus.

The format of chapters differs slightly depending on the stage of publication. In general, complex methods are presented at the end of the chapter, and simple methods are presented as they are used. Mathematical derivations and heuristic arguments are generally found in the relevant appendices.

Chapter 2

Phenotypic leverage with shared controls

2.1 Introduction

GWAS have enabled identification of genetic variants associated with a wide range of complex phenotypes, but in many cases these variants explain only a proportion of the known heritability [Hirschhorn and Daly, 2005]. It may be possible to improve variant discovery by re-analysis of existing data [Yang et al., 2011]. One promising strategy is to co-analyse GWAS results from similar phenotypes to exploit potential similarities in genetic aetiology. This has been attempted using several different methods [Cotsapas et al., 2011, Smyth et al., 2008, Ferkingstad et al., 2008].

A technique for improved discovery of disease variants using pleiotropy between pairs of diseases has been successfully developed and applied by Andreasson et al [Andreassen et al., 2013, Andreassen et al., 2014, Andreasson et al., 2014]. The technique extends the empirical Bayesian false discovery rate [Efron and Tibshirani, 2002] to a two-phenotype scenario, in which association with one phenotype is tested conditional on varying degrees of association with another. I denote the phenotype for which association is being tested as the 'principal phenotype' and the other as the 'conditional phenotype'.

By successively restricting attention to SNPs with a given strength of association in the conditional phenotype, the number of parallel tests to perform for association with the principal phenotype is reduced. If the two phenotypes share common associated variants, this restriction will retain disease-associated SNPs at a higher rate than null SNPs, resulting in a higher proportion of disease-associated SNPs in the restricted group than in the whole. The 'conditional false discovery rate' (cFDR), defined as the probability that a SNP is not associated the principal phenotype given its p values for the principal and conditional phenotypes are below some thresholds, exploits this effect. By computing cFDR values for

schizophrenia conditioned on bipolar disorder and vice versa, Andreasson et al [Andreassen et al., 2013] identified multiple previously undiscovered loci for both. In a separate study computing cFDRs for hypertension conditioned on 12 related traits [Andreassen et al., 2014], 42 new loci associated with hypertension were reported. These constituted considerable improvement on existing results using single GWAS, although using a liberal threshold of estimated $cFDR \leq 0.01$.

A major disadvantage of the algorithm developed and used by Andreasson et al is the requirement that control groups for the two GWAS be distinct, in order to ensure that observed effect sizes are uncorrelated at null SNPs. This requires splitting a pool of potential controls between studies, with the summary statistics for each GWAS computed from only the controls allocated to that study. This may be impractical as it requires access to raw genotype data. More importantly, accuracy of effect size estimates improves with larger control groups, and consequently splitting controls in this way weakens the effect size estimates for individual studies. For this reason, many researchers employ a study design in which controls are pooled into a large group; for example, the Wellcome Trust Case Control and ImmunoChip consortia [The Wellcome Trust Case Control Consortium, 2007, Cortes and Brown, 2011].

In this chapter I extend the cFDR approach to studies with overlapping control groups, exploiting an approach developed by Zaykin et al, following Lin et al [Zaykin and Kozbur, 2010, Lin and Sullivan, 2009] to adjust for the effect of shared controls. This allows the strongest available estimates of effect sizes to be used for calculation, and consequently strengthens the power of the technique. My method retains the ability of cFDR values to be computed from summary statistics alone, without the need to recalculate effect sizes after re-allocating controls. I demonstrate the improvement arising from sharing controls using data from a GWAS on T1D.

I also identify a previously un-discussed difficulty with the cFDR method in general potentially leading to a falsely low estimate of the false discovery rate amongst SNPs declared non-null. Multiple overlapping sets of SNPs may be defined, each of which has $cFDR \leq \alpha$. However, the union of these sets does not necessarily have a false-discovery rate less than α, and in general it is higher. An implication of this is that if we declare non-null all SNPs for which estimated cFDR is less than α, the overall false-discovery rate amongst SNPs declared non-null is greater than α. I describe an approximate upper bound on the false discovery rate amongst such SNPs based on areas of regions of the unit square.

I apply the method to summary SNP association statistics for ten variably phenotypically distinct autoimmune diseases: type 1 diabetes (T1D) [Onengut-Gumuscu et al., 2014], autoimmune thyroid disease (ATD) [Cooper et al., 2012], coeliac disease (CEL) [Trynka

et al., 2012], multiple sclerosis (MS) [Beecham et al., 2013], narcolepsy (NAR) [Faraco et al., 2013], primary biliary cirrhosis (PBC) [Liu et al., 2012], psoriasis (PS) [Tsoi et al., 2012], rheumatoid arthritis (RA) [Eyre et al., 2012], ulcerative colitis [Jostins et al., 2012], and Crohn's disease [Jostins et al., 2012]. All were genotyped using a common SNP array: the ImmunoChip, designed to provide dense genotype coverage of regions associated with autoimmune disease. Many autoimmune traits are known to have significant heritability, much of which remains unexplained [Cotsapas and Hafler, 2013]. I hypothesised that my method could improve detection of disease-associated variants in these diseases without the need for distinct control groups.

2.2 Results

2.2.1 Overview of method

Assume that each of the set of variants in question is either null ($H_0^{(i)}$) or non-null for association with the principal phenotype i. I consider observed p-values for phenotype i as observations of a random variable P_i with $P_i|H_0^{(i)} \sim U(0,1)$. Given a p-value threshold α, the positive false-discovery rate (equation 1.12 in chapter 1) is equivalent to $Pr(H_0|P_i < \alpha)$ under reasonable assumptions [Storey, 2002]. For each variant, we associate a test statistic termed the 'unconditional' false discovery rate on the basis of its observed p-value p_i as $uFDR(p_i) = Pr(H_0^{(i)}|P_i < p_i)$. This can be interpreted as 'the probability that a random SNP with $P_i < p_i$ is null for phenotype i'. The estimate of this quantity (identical to the left-hand side of inequality 1.11 in chapter 1) is denoted $\widehat{uFDR}(p_i)$.

The conditional false discovery rate (cFDR) [Andreassen et al., 2013, Andreasson et al., 2014] is the probability that a random SNP is null for a phenotype i given that the observed p values at that SNP for phenotypes i and j are less than (p_i, p_j); that is, $Pr(H_0^{(i)}|P_i \leq p_i, P_j \leq p_j)$, where $H_0^{(i)}$ is the null hypothesis that the SNP is not associated with phenotype i. This quantity is denoted $cFDR(p_i|p_j)$ and the estimate as $\widehat{cFDR(p_i|p_j)}$. As above, phenotype i is termed the 'principal phenotype' and phenotype j the 'conditional phenotype'.

I first apply genomic control to allow the assumption that, globally, P values for null SNPs are uniformly distributed on $[0,1]$. I compute an estimate $\widehat{cFDR}(p_i|p_j)$ of the $cFDR$ in a similar manner to that proposed by Andreasson et al, but incorporating expected non-uniformity in the distribution of P_i due to the sharing of controls. As $\widehat{uFDR}(p_i)$ is monotonically related to p_i, I set a significance cutoff for $\widehat{uFDR}(p_i)$ at the maximum value of $\widehat{uFDR}(p_i)$ with $p_i < 5 \times 10^{-8}$. Correspondingly, I set a significance cutoff for $\widehat{cFDR}(p_i, p_j)$ at the maximum

$\widehat{cFDR}(p_i, p_j)$ with $p_i < 5 \times 10^{-8}$ [1] Implementation of these steps in R is available from https://github.com/jamesliley/cFDR-common-controls.

2.2.2 Procedures for uFDR and cFDR

Assume that the p-values for a phenotype i across all SNPs are instances of a random variable P_i. If p_i is an instance of this random variable corresponding to a SNP of interest, the unconditional false discovery rate $uFDR(p_i)$ is defined as

$$\begin{aligned} uFDR(p_i) &= Pr(H_0^{(i)} | P_i \leq p_i) \\ &= Pr\left(H_0^{(i)}\right) \frac{Pr(P_i \leq p_i | H_0^{(i)})}{Pr(P_i \leq p_i)} \\ &= Pr\left(H_0^{(i)}\right) \frac{p_i}{Pr(P_i \leq p_i)} \end{aligned}$$

where $H_0^{(i)}$ is the null hypothesis that the SNP of interest is not associated with phenotype i. Given a set of observed p values $\{p_i^1, p_i^2 ... p_i^N\}$ for a phenotype i at N different SNPs, and an observed p value p_i for a SNP of interest, this quantity is estimated as

$$\begin{aligned} \widehat{uFDR}(p_i) &= \frac{p_i}{\#(\text{p values } p_i^k \text{ with } p_i^k \leq p_i)/N} \\ &= \frac{\text{Expected quantile of } p_i \text{ under } H_0^{(i)}}{\text{Observed quantile of } p_i} \end{aligned} \tag{2.1}$$

Because of the approximation $Pr(H_0^i) = 1$, the estimate $\widehat{uFDR}$ is a an upwards-biased estimate of $uFDR$; that is, its expected value is greater than the true $uFDR$, making it a conservative estimate.

I compute the quantity (2.1) for each SNP at each phenotype, declaring any SNP for which $\widehat{uFDR}(p_i) \leq \alpha$ as non-null for phenotype i. Defining V as the number of SNPs falsely declared non-null, R as the total number of SNPs declared non-null, and $Q = V/R$, a theorem of Benjamini and Hochberg [Benjamini and Hochberg, 1995] shows the false discovery rate $E(Q)$ among SNPs with $\widehat{uFDR} \leq \alpha$ is less than α.

[1] This cutoff is liberal relative to the genome-wide significance threshold, but is conservative relative to that previously used in cFDR-based studies. In subsequent work discussed in chapter 3, I argue that the threshold used here is too liberal and propose a more appropriate one. However, the choice of threshold is a matter of preference rather than necessity, and in this chapter I have retained the choice of threshold used in the relevant publication.

The *cFDR* constitutes a natural extension of this idea. I assume that the p-values for two phenotypes i and j across all SNPs are instances of a pair of random variables P_i, P_j. If p_i and p_j are instances of these variables corresponding to a SNP of interest then the conditional false discovery rate $cFDR$ is defined for the set of SNPs with p values for each phenotype less than or equal to those at this SNP (as per Andreasson et al [Andreassen et al., 2013]) as

$$\begin{aligned} cFDR(p_i|p_j) &= Pr(H_0^{(i)}|P_i \leq p_i, P_j \leq p_j) \\ &= Pr(H_0^{(i)}|P_j \leq p_j)\frac{Pr(P_i \leq p_i|P_j \leq p_j, H_0^{(i)})}{Pr(P_i \leq p_i|P_j \leq p_j)} \end{aligned} \tag{2.2}$$

The estimation of this quantity proceeds in a similar way to *uFDR*. Given a set of observed p value pairs $\{(p_i^1, p_j^1), (p_i^2, p_j^2)...(p_i^N, p_j^N)\}$ for two phenotypes i and j at N different SNPs, and an observed p value pair (p_i, p_j) for a SNP of interest, define N_1 as the number of p value pairs with $P_j \leq p_j$, and estimate the $cFDR$ as

$$\begin{aligned} \widehat{cFDR}(p_i|p_j) &= \frac{Pr(P_i \leq p_i|P_j \leq p_j, H_0^{(i)})}{\#(\text{pairs } (p_i^k, p_j^k) \in (P_i, P_j) \text{ with } p_i^k \leq p_i \text{ and } p_j^k \leq p_j)/N_1} \\ &= \frac{\text{Expected quantile of } p_i \text{ under } H_0^{(i)} \text{ amongst } p_i^k \text{ with } k \text{ satisfying } p_j^k \leq p_j}{\text{Observed quantile of } p_i \text{ amongst } p_i^k \text{ with } k \text{ satisfying } p_j^k \leq p_j} \end{aligned} \tag{2.3}$$

Again, this estimate is conservative, due to the approximation $Pr(H_0^{(i)}|P_j \leq p_j) = 1$.

I compute the quantity (2.2) for each SNP at each pair of phenotypes, declaring any SNP for which $\widehat{cFDR}(p_i|p_j) \leq \alpha'$ as non-null for phenotype i. However, as noted earlier, this does not guarantee that the expected false discovery rate amongst such SNPs is less than α'. I show that the FDR is approximately controlled at a higher level dependent on the region of the unit square defined by rectangles for which $\widehat{cFDR}(x|y) \leq \alpha'$.

My method here diverges from the original method proposed by Andreasson et al, in the use of the expected quantile $Pr(P_i \leq p_i|P_j \leq p_j, H_0^{(i)})$ in place of the p-value p_i. If studies share no controls, it can be reasonably assumed that, for a SNP which is null for phenotype i, the p values (p_i, p_j) are independent, so $p_i' = Pr(P_i \leq p_i|P_j \leq p_j, H_0^{(i)}) = p_i$. This is the approach taken by Andreasson et al [Andreassen et al., 2013]. I propose a method for computing p_i' when controls are shared between studies, and the independence assumption above is not valid.

My approach is to compute the related quantity $Pr(P_i \leq p_i|P_j \leq p_j, H_0^{(i)}, H_\eta^{(j)})$, where η is the (unobserved) effect size we would observe for a given SNP for phenotype j if the observed MAFs agreed exactly with the population MAFs for that SNP, and $H_\eta^{(j)}$ is the

hypothesis that $Z_j \sim N(\eta, 1)$ for that SNP. This quantity can be thought of as the 'expected quantile' of p_i; that is, the proportion of p values we expect to be less than p_i.

2.2.3 Sharing of control subjects

If no controls are shared between studies, it is reasonable to assume that observed effect sizes for the two phenotypes are independent under a null hypothesis for the principal phenotype. This implies that the expected quantile of a given SNP's p value for the principal phenotype is simply the p value itself regardless of its p value for the conditional phenotype. However, when control samples are shared, this assumption is invalid. Shared controls induce a positive correlation on estimated effect sizes for the principal and conditional phenotype [Zaykin and Kozbur, 2010, Lin and Sullivan, 2009], meaning that when attention is restricted to SNPs with a given degree of association with the conditional phenotype, the p values for the principal phenotype will be falsely low; that is, the probability $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ will not in general be equal to p_i $(= Pr(P_i \leq p_i | H_0^{(i)}))$; in fact it will usually be higher.

When controls are shared, the distribution of p values for the principal phenotype given p values for the conditional phenotype depends on the underlying effect of each SNP on the conditional phenotype. For any given SNP, this underlying effect size, which I denote η, is not known. However, across all SNPs, η may be considered to be realisations of a random variable H whose distribution is mirrored by the distribution of observed effect sizes for the conditional phenotype. By integrating over this unknown true effect size for the conditional phenotype, allowance can be made for shared controls, and the 'expected quantile' of a p value for the principal phenotype, defined as $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$, can be calculated (section 2.4).

Assume that H has a mixture distribution defined by two parameters (π_0, σ^2), such that $H = 0$ with probability π_0 and $H \sim N(0, \sigma^2)$ with probability $1 - \pi_0$. The parameters (π_0, σ^2) are estimated from the observed distribution of effect sizes for the conditional phenotype. In order to show the effect of the p value adjustment, I simulated p values for 20,000 SNPs for a principal and conditional phenotype, with controls shared between simulated studies. All SNPs were null for the principal phenotype, and were variably null or non-null for the conditional phenotype with probability 0.9, 0.1 respectively. Z scores at non-null SNPs for the conditional phenotype were distributed as $N(0, \sigma^2)$, as per our assumption. I used a value of $\sigma = 3$, which was similar to the values of σ in real data estimated by our E-M algorithm.

Consider the set of simulated SNPs with p values for the conditional phenotype less than 0.05 (Figure 2.1). In the absence of shared controls, the distribution of p_i amongst

this set is expected to be uniform, and hence in figure 2.1 the black dots would be expected to lie along the x-y line. However, with shared controls the principal p values are biased downward in this set (black dots, figure 2.1). Our computed expected quantile (blue dots) agrees closely with the observed quantile. In a sense, this constitutes 'adjusting' the p values for the principal phenotype so that the expected distribution is uniform under the null hypothesis. Software to generate this simulation is available at https://github.com/jamesliley/cFDR-common-controls.

The formula for adjusting p-values can easily be adapted to arbitrary distributions of H at the cost of increased computational time, but the form of the distribution of H is not generally known. I show in appendix A, section A.1.1 that, even for distributions of H which differ markedly from our assumption of normality, the error in the estimate is not large, and generally translates to a negligible difference in the set of SNPs declared non-null using the $\widehat{cFDR}$ method. While my assumption of normality has the potential to be anti-conservative if the true distribution of H is bimodal, non-parametric estimates for distributions of effect sizes suggest they have a uni-modal distribution centred on zero [Park et al., 2010]. Reassuringly, the assumption of normality is conservative if H has heavier tails than a normal.

2.2.4 Comparison to split control approach

I compared Andreasson's approach to SNP discovery which advocated splitting controls into non-overlapping subsets to my extended shared-control approach using a type 1 diabetes dataset with a total of 12,175 cases and 15,171 controls. Controls and cases were each split into two sets (control sets had size 7,585 and 7,586, cases 6087 and 6088). 'Split' p values were computed using one set of controls and one set of cases and corresponding 'shared' p values were computed using the complete set of controls. As expected, more shared p values reached genome-wide significance than did split p values (Figure 2.2).

I computed $cFDR$ values by labelling one set of cases 'conditional' and the other 'principal' using the split-control p values using Andreasson et al's approach and using the shared control p values using our method. For reference, I compared these to a naive application of Andreasson et al's method on the shared-control p values (Figure 2.2, lower panel). More SNPs can be declared significant according to $cFDR$ using the shared-control than split-control approach at all reasonable thresholds, and naive application of Andreasson et al's approach to shared-control p values again increases the number declared significant.

Because the quantity $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ is systematically underestimated when using this naive method (by assuming it is equal to p_i) as shown in appendix A, section A.1.3,

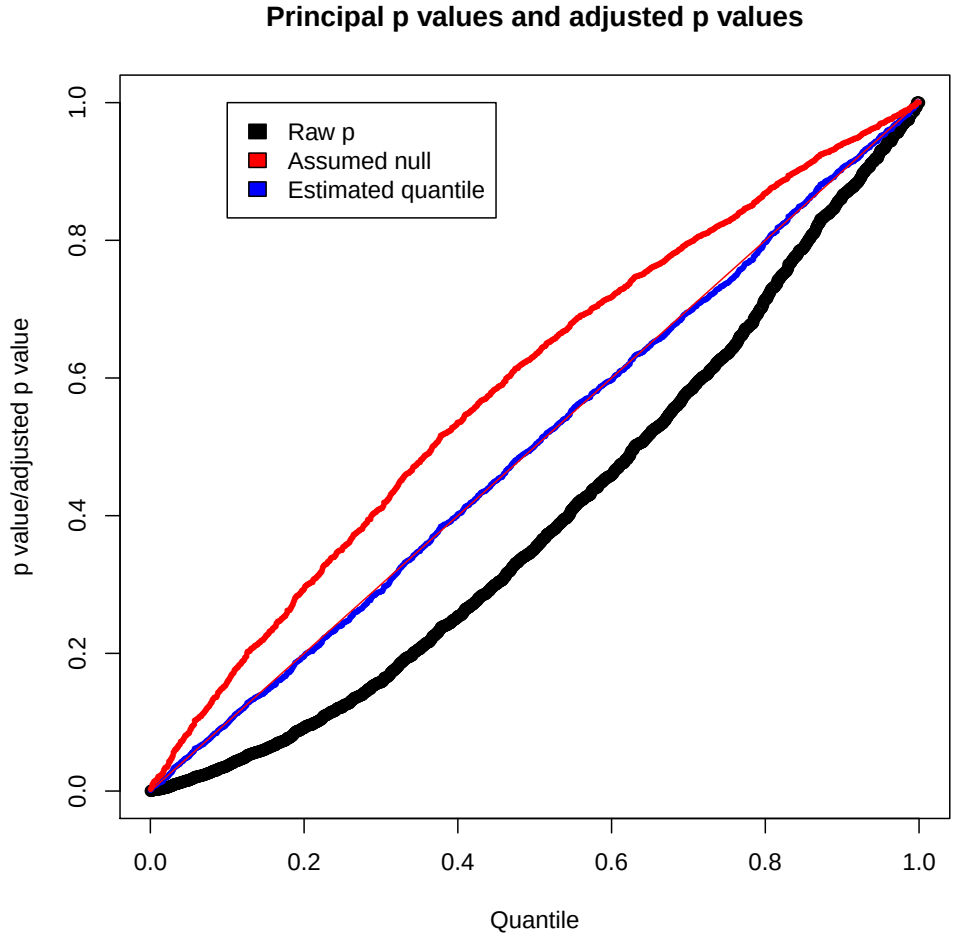


Fig. 2.1 Correction of cFDR for shared controls. Simulation of GWAS summary statistics for 20000 SNPs, all null for phenotype i and variably null or non-null for phenotype j, with association tested using a shared control group. Black dots show p values for phenotype i at all SNPs with p value for phenotype j less than 0.05, with evident downward bias. Blue dots show the adjustment to expected quantile of p values. The red dots show the expected quantile we would compute if we were to assume incorrectly that all SNPs were null for the conditional phenotype i. This quantity overestimates the true quantile.

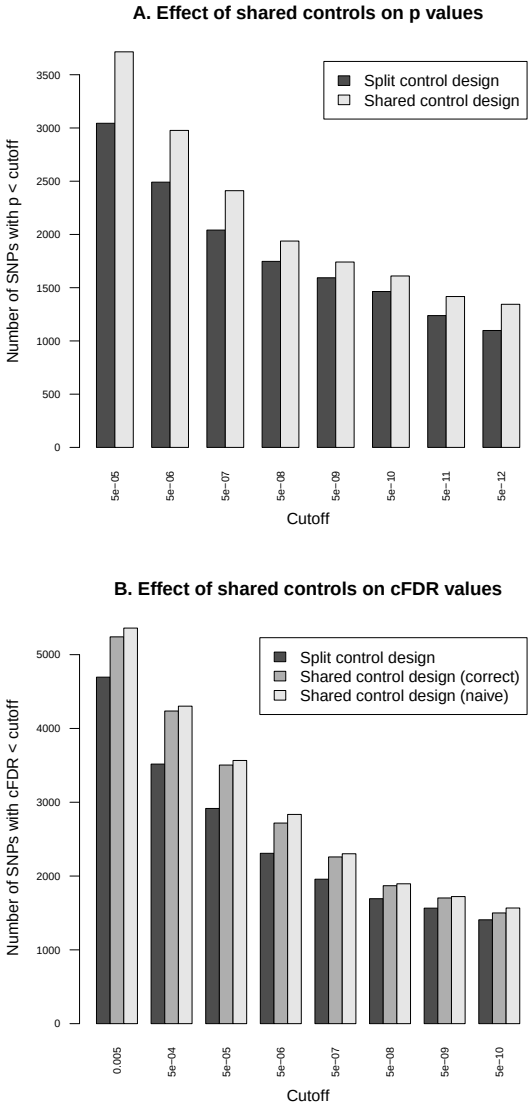


Fig. 2.2 Validation of the shared-control approach and p-value adjustment due to shared controls. Panel A shows the effect of splitting controls on power to detect association. The number of SNPs with p values less than a given cutoff are shown for split-control and shared-control approaches. For all p-value cutoffs, fewer SNPs reach significance when using a split-control design. Panel B shows the number of SNPs with $\widehat{cFDR}$ values less than a given cutoff using the existing method on a split-control design, our extended method on a shared control design with the adjustment for shared controls, or using the split-control approach naively on the shared-control design; that is, assuming incorrectly that $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) = p_i$. The second figure shows that failing to correctly calculate $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) = p_i$ leads to a subtle increase in the number of SNPs declared non-null at all cutoffs, due to the incorrect underestimation of $\widehat{cFDR}$.

it leads to a falsely low $\widehat{cFDR}$. The increase in observed number of SNPs declared significant when using the naive method shows that it can indeed lead to false discoveries.

For principal phenotype p values in the range 5×10^{-6} - 5×10^{-8} - effectively the region from which 'new' SNPs may be discovered by cFDR rather than p value alone - the naive $cFDR$ is frequently underestimated by 2-3 fold (appendix A, figure A.5, left panel). For lower p values, the naive $cFDR$ may underestimate by hundreds- or thousand-fold, with the potential fold underestimation increasing with decreasing p value (appendix A, figure A.5, right panel). Because of the relatively high ratio of number of controls to number of cases, the correlation between effect sizes is lower in this constructed case (c. 0.22) than between most phenotypes in our study (c. 0.5). The underestimation of $cFDR$ using the 'naive' method worsens with higher correlation, so we would expect that the fold-underestimate we see here is less severe than that which would be observed if applying this to other studies.

2.2.5 False discovery rate

An important consideration in any GWAS procedure is control of the type-1 error rate. This can be done in several ways, as discussed in chapter 1, section 1.2.4. A major advantage of the $\widehat{uFDR}$ statistic is the control it allows over the FDR by the Benjamini-Hochberg procedure: if $H_0^{(i)}$ is rejected for all SNPs with $\widehat{uFDR}$ less a given threshold α, the FDR amongst these SNPs is bounded above by α.

This is not true for the $\widehat{cFDR}$. This can be seen most easily by considering an extreme case in which all non-null SNPs for phenotype i have $P_i = P_j = 0$ (figure 2.3). Assume no controls are shared and no SNPs are non-null for phenotype j and null for phenotype i, so $P_i|H_0^{(i)}$ and $P_j|H_0^{(i)}$ have independent $U(0,1)$ distributions. Assume also that the total number of SNPs N is large but a fixed positive proportion of SNPs $\pi_1 = 1 - \pi_0$ are non-null. For fixed p_i, p_j, the number of non-null SNPs with $Pr(P_i < p_i, P_j < p_j) = p_i p_j$ so as $N \to \infty$

$$\widehat{cFDR}(p_i|p_j) \to_{A.S} p_i \frac{\pi_0 p_j + \pi_1}{\pi_0 p_i p_j + \pi_1} \tag{2.4}$$

For a region X of the unit square, define (as per Storey et al [Storey et al., 2003]) the random variables $V(X)$, $S(X)$ and $R(X)$ as the number of null SNPs, non-null SNPs and the total number of SNPs for which observed p-values fall in X. Defining X as the rectangle with corners $(0,0)$, $(p_i,0)$, $(0,p_j)$, (p_i,p_j), as $N \to \infty$, we have

$$V(X)/R(X) \to_{A.S} \frac{p_i p_j \pi_0}{p_i p_j \pi_0 + \pi_1} < p_i \frac{\pi_0 p_j + \pi_1}{\pi_0 p_i p_j + \pi_1} \tag{2.5}$$

so for large enough N the $\widehat{cFDR}$ overestimates the FDR for the rejection region X. Indeed, if p'_j is fixed and $H_0^{(i)}$ is rejected whenever $P_j < p'_j$ and $\widehat{cFDR}(p_i|p'_j) < \alpha$, the FDR is controlled at α by the main result of Benjamini and Hochberg [Benjamini and Hochberg, 1995]. However, as $N \to \infty$ the region for which $\widehat{cFDR} < \alpha$ tends to the region L defined by

$$\begin{aligned}(p_i, p_j): \quad & p_i \frac{\pi_0 p_j + \pi_1}{\pi_0 p_i p_j + \pi_1} < \alpha \\ & \Leftrightarrow p_j < \frac{\pi_1}{\pi_0(1-\alpha)} \left(\frac{\alpha}{p_i} - 1 \right)\end{aligned} \tag{2.6}$$

and, by integrating over this region, we have

$$\begin{aligned}V(L) &\to_{A.S} N \frac{\pi_1 \alpha}{(1-\alpha)\pi_0} \log\left(\frac{1-\alpha\pi_0}{\pi_1} \right) \\ R(L) &\to_{A.S} N \left(\pi_1 + \frac{\pi_1 \alpha}{(1-\alpha)\pi_0} \log\left(\frac{1-\alpha\pi_0}{\pi_1} \right) \right) \\ \frac{V(L)}{R(L)} &\to_{A.S} \frac{\alpha \log\left(\frac{1-\alpha\pi_0}{1-\pi_0} \right)}{\pi_0(1-\alpha) + \alpha \log\left(\frac{1-\alpha\pi_0}{1-\pi_0} \right)}\end{aligned} \tag{2.7}$$

which is not generally $< \alpha$ (in fact, it is usually substantially $> \alpha$, and may be larger than α by an arbitrarily large factor) so the procedure in which $H_0^{(i)}$ is rejected whenever $\widehat{cFDR} < \alpha$ does not asymptotically control the FDR at α (or at $k\alpha$ for any fixed k).

This is demonstrated in figure 2.3. For any preset p'_j, the region defined by $P_j < p'_j$, $\widehat{cFDR}(p_i|p'_j) < \alpha$ is a rectangle with a corner on the boundary of L. Since any such rectangle M contains all non-null SNPs (and thus as many as L itself) and fewer null SNPs, each rectangle has a lower value of V/R than does L. Intuitively, the Benjamini-Hochberg result only guarantees FDR control over regions like M, not over regions like L.

In the original method [Andreassen et al., 2013], SNPs were declared significant if they were contained within any rectangular regions with a $\widehat{cFDR}$ value of less than 0.01. Our reasoning demonstrates that the false-discovery rate of this procedure was likely to have been considerably higher than 0.01.

It is a difficult problem to find a bound on FDR for the procedure in which SNPs are rejected if $\widehat{cFDR} < \alpha$. However, given a rejection region L for p-value pairs (p_i, p_j), we can compute a bound on the positive FDR of the region L, $Pr(H_0^{(i)}|(p_i, p_j) \in L)$; that is, the FDR if we were to repeatedly re-generate and re-test P_i, P_j values while retaining the same rejection region L.

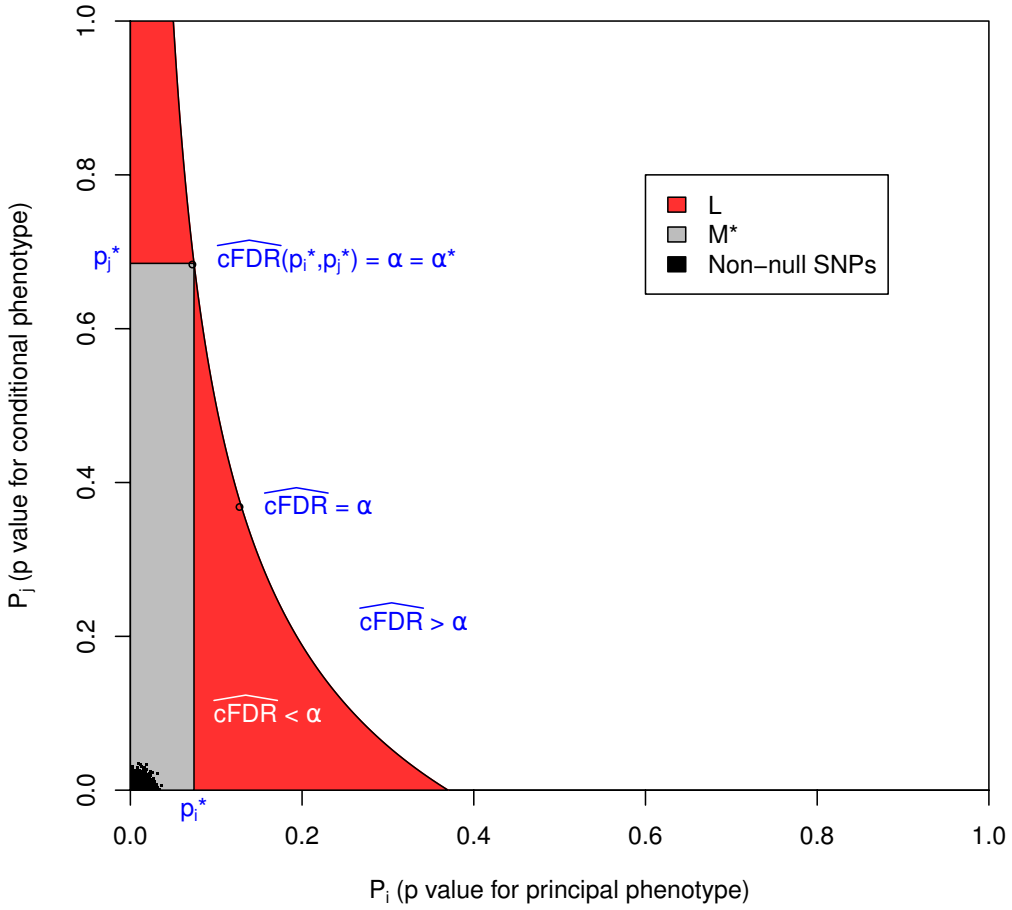


Fig. 2.3 FDR control on SNPs reaching cFDR threshold. L is the locus of a set of points with $\widehat{cFDR} = \alpha$, and M^* is a rectangle within L. If all the non-null SNPs were concentrated in the lower left corner, then the number of non-null SNPs in L would be equal to that in any individual rectangle with vertices at the origin and on L, but the number of null SNPs would be greater, meaning that $V(L)/R(L)$ would be greater than $V(M)/R(M)$, and generally greater than α. The value $E\{V(L)/R(L)|R(L) > 0\}$ is asymptotically bounded above by $\alpha^* v(L)/v(M)$, where $v(L)$ and $v(M^*)$ are the expected number of null SNPs in L and M^* respectively (appendix A, section A.1.2).

From Storey et al [Storey et al., 2003], we assume that we are performing N identical hypothesis tests to generate bivariate test statistics (P_{ik}, P_{jk}), $k = 1..N$. Denote $H_{0k}^{(i)}$ as the null hypothesis for test k and assume that $H_{0k}^{(i)}$ has a Bernoulli distribution with parameter $1-\pi_0$, and that the tri-variate random variables $(P_{ik}, P_{jk}, H_{0k}^{(i)})$ are independent and identically distributed with $(P_{ik}, P_{jk})|H_{0k}^{(i)} \sim (1-H_{0k}^{(i)})F_0 + H_{0k}^{(i)}F_1$. Further assume that F_0 is known, and for each region X denote

$$v(X) = \int_X dF_0 \tag{2.8}$$

Set M as a rectangle contained within L with vertices $(0,0)$, $(p_i^*, 0)$, $(0, p_j^*)$ and (p_i^*, p_j^*). Now, twice using the result that $E\left(\frac{V(X)}{R(X)}\middle|R(X) > 0\right) = \frac{E\{V(X)\}}{E\{R(L)\}} = Pr(H_0^{(i)}|(p_i, p_j) \in X)$ [Storey et al., 2003], we have

$$\begin{aligned}
Pr(H_0^{(i)}|(p_i, p_j) \in L) &= E\left(\frac{V(L)}{R(L)}\middle|R(L) > 0\right) \\
&= \frac{E\{V(L)\}}{E\{R(L)\}} \\
&= \frac{\pi_0 v(L)}{\pi_0 v(L) + (1-\pi_0)E\{S(L)\}} \\
&< \frac{\pi_0 v(L)}{\pi_0 v(M) + (1-\pi_0)E\{S(M)\}} \\
&= \frac{v(L)}{v(M)} \frac{\pi_0 v(M)}{\pi_0 v(M) + (1-\pi_0)E\{S(M)\}} \\
&= \frac{v(L)}{v(M)} \frac{E\{V(M)\}}{E\{R(M)\}} \\
&= \frac{v(L)}{v(M)} E\left(\frac{V(M)}{R(M)}\middle|R(M) > 0\right) \\
&= \frac{v(L)}{v(M)} Pr(H_0^{(i)}|(p_i, p_j) \in M)
\end{aligned} \tag{2.9}$$

The value $\alpha^* = \widehat{cFDR}(p_i^*|p_j^*)$ is an estimate for $Pr(H_0^{(i)}|(p_i, p_j) \in M)$ which is almost surely conservative as $N \to \infty$ (under reasonable conditions on F_0, F_1). Thus $\frac{v(L)}{v(M)}\alpha^*$ provides an asymptotic bound for $Pr(H_0^{(i)}|(p_i, p_j) \in L)$, and a rough bound for the FDR for the $\widehat{cFDR}$ procedure.

By the definition of L in our procedure, $\alpha = \alpha^*$, and it is desirable to choose the rectangle M^* in L which minimises $\frac{v(L)}{v(M)}$. Importantly, this bound is not strict for finite N, since the region L is not independent of the observations of P_i, P_j. It will, however, hold almost surely for fixed α, π_0, F_0, F_1 as $N \to \infty$ (since the rejection region will converge almost surely to a

limit L^∞, in a similar way to equation 2.6, and $\alpha v(L^\infty)/v(M^*)$ is almost surely asymptotically conservative) under reasonable assumptions on F_0, G_0. For the remainder of this thesis, I will refer to this as a 'bound', with these caveats assumed.

The estimation of F_0 and hence $v(X)$ with shared controls and $P_i|H_0^{(i)} \nsim U(0,1)$ is discussed in appendix A, section A.1.2.

2.2.6 Application to ten immune mediated diseases

I obtained summary statistics in the form of p values for ten immune mediated diseases from ImmunoBase (`www.immunobase.org`, accessed 19/3/14). For each pair of diseases, the number of shared controls was estimated according to the description of the control samples in each paper. The numbers of cases, controls and our estimated numbers of shared controls for each study are shown in Table 2.1. Uniform quality control criteria were applied to all SNPs, and the MHC region, which exhibits both strong LD and strong association with immune mediated diseases was excluded. P values were corrected within each trait for genomic inflation using a standard algorithm [Devlin et al., 2001] applied to SNPs included on the ImmunoChip to replicate a GWAS study of reading and maths ability (Steve Eyre and Cathryn Lewis, personal communication), unlikely to be related to any immune mediated disease studied here.

P values for each principal phenotype were adjusted to p' as described above in order to account for the effect of shared controls. For each ordered pair of phenotypes, a Q-Q plot was generated as per Andreasson et al [Andreassen et al., 2013]. A Q-Q plot is a graph of the observed distribution of a random variable against the expected distribution. I overlaid Q-Q plots for $\log_{10}(p')$ values for the principal phenotype for subsets of SNPs exhibiting successively smaller p values for the conditional phenotype. Figure 2.4 shows QQ plots for T1D conditional on RA and PSO; plots for all other pairwise comparisons may be found in appendix A, figure A.8 onwards. Notably, if lines shift further left with more stringent cutoffs on association with the conditional phenotype, then SNPs which are associated with the conditional phenotype are more likely to be associated with the principal phenotype, indicating pleiotropic effects of SNPs on the two phenotypes. In many cases, the Q-Q plots demonstrate considerable leftward shift with conditioning on association with a second disease, and we see strong evidence for pleiotropy for T1D conditioned on RA and little or no evidence for pleiotropy for T1D conditioned on PSO.

I estimated the unconditional and conditional false discovery rates, $\widehat{uFDR}(p_i)$ and $\widehat{cFDR}(p_i|p_j)$, at each SNP for each phenotype and each ordered pair of phenotypes re-

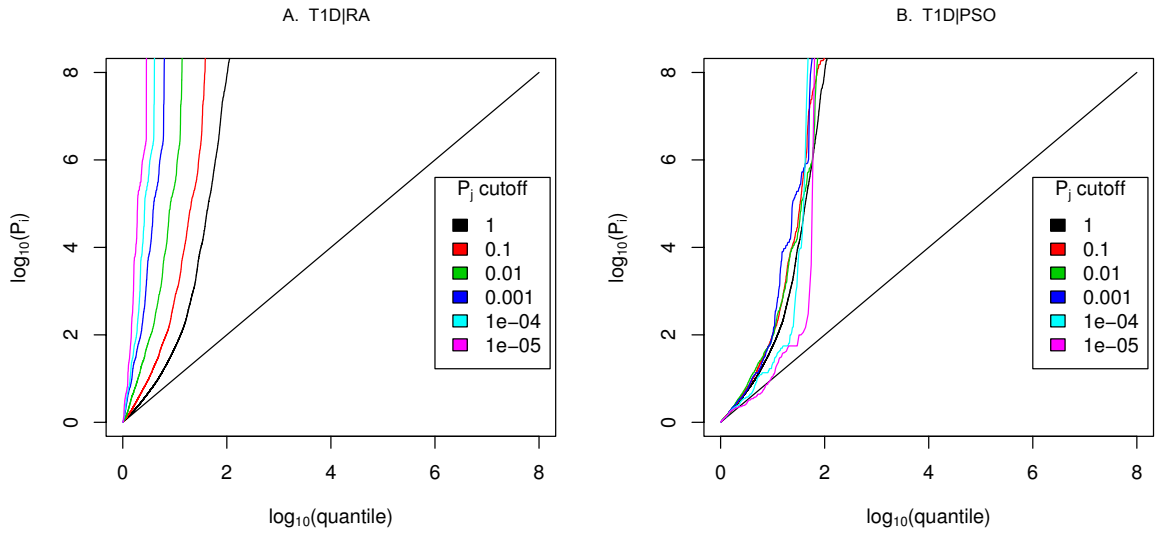


Fig. 2.4 Q-Q plots for T1D conditional on RA (Panel A) and PSO (Panel B). Y axes show $\log_{10}(p'_{T1D})$; X axes show log quantile (rank) of p values in various sets of SNPs. The degree of leftward shift of a black point from the diagonal is monotonic with the unconditional FDR of that p value for the principal phenotype, and the degree of leftward shift of a coloured point is monotonic with the conditional FDR of the p value for the principal phenotype and the p-cutoff corresponding to the colour for the conditional phenotype. As expected, a leftward shift is seen even for the unconditional Q-Q plots (black line) owing to the use of the ImmunoChip, which focuses on potential autoimmune-associated regions. Each colour corresponds to the Q-Q plot for p_{T1D} amongst a subset of SNPs with p_{RA} or p_{PSO} less than the indicated cutoff. P values for T1D are adjusted for the effect of shared controls between studies. A leftward shift with decreasing p_{RA} or p_{PSO} cutoff indicates that SNPs which are associated with the conditional phenotype (RA or PSO) are more likely to be associated with the principal phenotype (T1D), presumably due to pleiotropic effects on phenotypes. Good enrichment is seen for T1D conditioning on RA (Panel A), and little or no enrichment conditioning on PSO (Panel B).

Table 2.1 Number of cases and controls for each study, and relevant references, together with our estimates of the number of controls shared between studies. P values for T1D are from a meta-analysis of case-control and TDT data, with effective numbers of cases computed as shown in the methods section

Disease		Controls	Cases
T1D	[Onengut-Gumuscu et al., 2014]	12175	15171
ATD	[Cooper et al., 2012]	9364	2733
CEL	[Trynka et al., 2012]	12228	12041
MS	[Beecham et al., 2013]	24091	14498
NAR	[Faraco et al., 2013]	10421	1886
PBC	[Liu et al., 2012]	8514	2861
PSO	[Tsoi et al., 2012]	22806	10588
RA	[Eyre et al., 2012]	15870	11475
UC	[Jostins et al., 2012]	15977	10920
CRO	[Jostins et al., 2012]	15977	14763

Estimated number of pairwise shared controls

	T1D	ATD	CEL	MS	NAR	PBC	PSO	RA	UC	CRO
T1D	-	9364	12228	8430	4289	8514	4822	8430	4020	4020
ATD		-	9364	8430	4289	8514	4822	8430	4020	4020
CEL			-	8430	4289	8514	4822	8430	4020	4020
MS				-	4289	8430	4822	8430	10102	10102
NAR					-	4289	4289	4289	4020	4020
PBC						-	4822	8430	4020	4020
PSO							-	4822	4020	4020
RA								-	4020	4020
UC									-	15977
CRO										-

spectively. Figure 2.5 shows $\widehat{cFDR}$ for T1D conditioned on RA. The advantage gained by $\widehat{cFDR}$ can be seen in the left-shift of the region in which a SNP can be declared significant (blue dots), corresponding to a higher p-value cutoff for significance for T1D among SNPs with low p values for RA. Indeed, if only SNPs with a p value for RA less than some threshold ζ are considered, a p value cutoff for significance for T1D is given by the leftmost border of the blue dots on the line $P_j = \zeta$.

The degree of leftward shift in the Q-Q plots clearly contains information about the degree of pleiotropy between diseases. I defined a statistic summarizing some aspects of this evidence for pleiotropy and used it to visualise the set of pairwise relationships between diseases as a network (Figure 2.6). The network encouragingly reflects several pathophysiological associations: UC is linked to CRO, and T1D to ATD. Strong linkage is

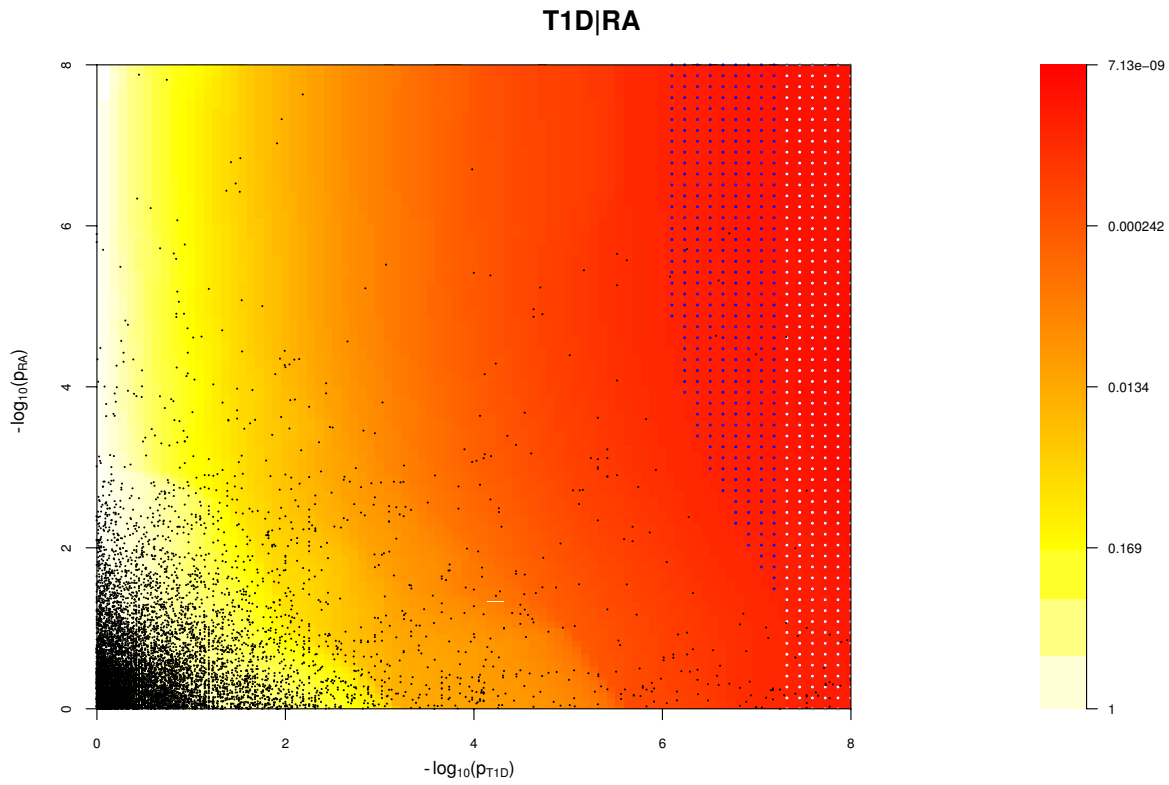


Fig. 2.5 Plot of $\widehat{cFDR}$ (red-yellow) for T1D conditioned on RA. White dots signify the region for which $\widehat{uFDR}$ is less than α corresponding to $p < 5 \times 10^{-8}$. Blue dots signify the region for which cFDR is less than the same α. Note the leftward shift of blue points and the general leftward shift of colours corresponding to an increased p-value threshold for association with T1D for SNPs with low p values for RA. Black dots show a random sample of the observed p value pairs.

also seen both ways for MS and PBC, and between T1D and RA, findings which can also be seen in the Q-Q plots (A, figure A.8 onwards). One way relationships suggest the presence of a larger total number of associated SNPs for the disease at the start of the arrow than at the end.

2.2.7 Discovery of novel associations

The numbers of SNPs deemed significant for each phenotype by analysis using unconditional and conditional approaches are shown in table 2.2, with details in A, table A.2 onwards. $\widehat{cFDR}$ allows certain SNPs with p values as high as 3×10^{-6} to be declared significant while controlling the false discovery rate at a relatively low value. Fifty-one of the 59 SNPs we identify uniquely through $\widehat{cFDR}$ have previously been reported to be associated with the relevant disease through use of alternative significance thresholds, other genomic control procedures, other GWAS or additional samples not genotyped by ImmunoChip, a useful verification of our technique. Eight of the SNPs we discover uniquely through cFDR were in regions not previously known to be associated with the corresponding disease (table 2.3). These will require replication in independent samples to be declared truly associated, but they contain some potentially interesting signals, such as an association for RA at SNP rs72928038 near existing MS, ATD and T1D associations in *BACH2*, a transcriptional regulator involved in transcription repression and activation by *MAFK* [Dubois et al., 2010]. The biological plausibility of new findings is somewhat expected, given the restricted coverage of the ImmunoChip to potentially immune-associated regions.

The SNP rs1034290 in region 1p13.1, which we found to be associated with PBC, is in intron three of *CD58*, which is a surface receptor involved in binding and activation of T-lymphocytes. The protective effect of the MS-associated allele is postulated to arise from upregulation of the transcription factor FOXP3 [De Jager et al., 2009] and the patterns of association in the region suggest the two diseases may share a causal variant here (http://www.immunobase.org).

2.3 Discussion

In this chapter, I extended a technique for computing conditional Bayesian False Discovery Rates to GWAS for independent diseases with shared control groups. This technique enables improved detection of disease-associated SNPs compared to conventional methods. By enabling larger control groups for each study, my method uses data more efficiently than in

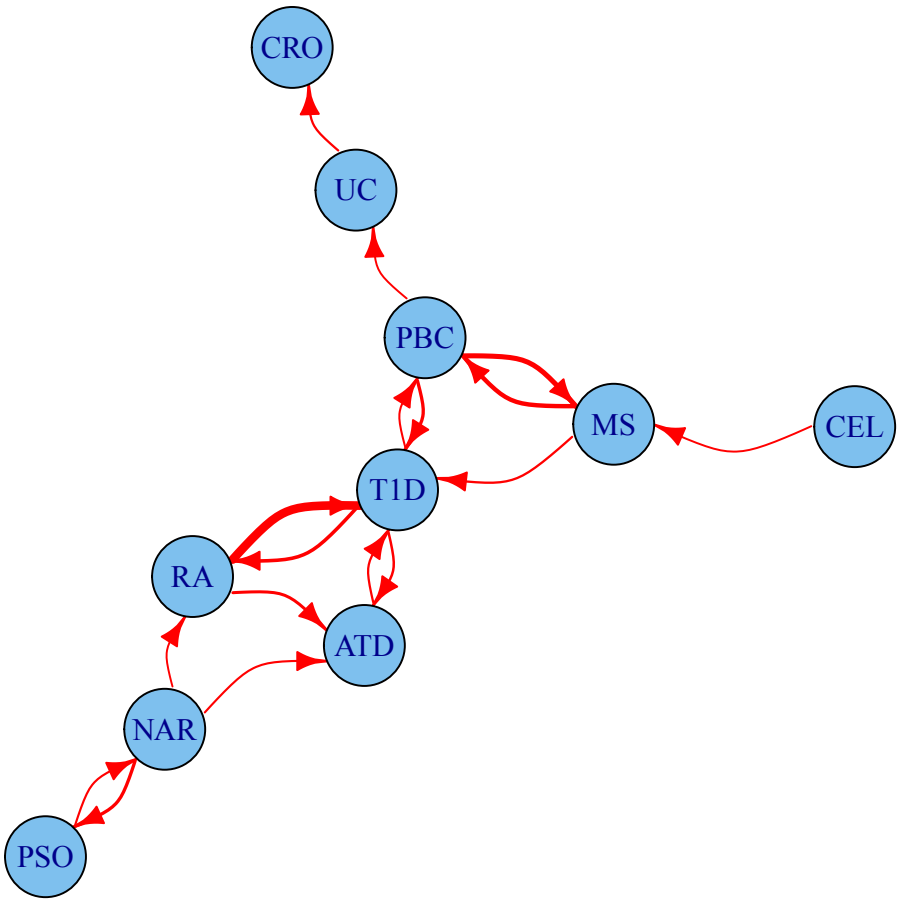


Fig. 2.6 Network of degree of pleiotropy between phenotypes (CRO: Crohn's disease; UC: Ulcerative colitis; PBC: primary biliary cirrhosis; MS: multiple sclerosis; CEL: coeliac disease; T1D: type 1 diabetes; RA: rheumatoid arthritis; ATD: autoimmune thyroid disease; NAR: narcolepsy; PSO: psoriasis). An arrow runs from vertex i to vertex j if and only if by conditioning on $p < 5 \times 10^{-6}$ for the conditional phenotype j we can increase the threshold for significance for the p value for the principal phenotype i from 5×10^{-8} to 4×10^{-7} or greater. Edges are thickened if the cutoff could be increased more than this. The threshold 4×10^{-7} was selected as the minimum value for which the network is weakly connected; that is, having an arrow to or from each edge.

Table 2.2 Number of SNPs with $p \leq 5 \times 10^{-8}$ after genomic found by analysis of the principal phenotype alone and the estimated equivalent FDR for this set of SNPs. Conditional analysis shows the number of additional SNPs found through conditional FDR analysis, and the upper bound for the FDR of all SNPs selected by this method, including adjustment for the multiple phenotypes conditioned upon. Finally, we summarise the performance of the cFDR approach by the FDR ratio - the ratio of the upper bound for the FDR for all SNPs selected by the cFDR approach, and the estimated FDR of all SNPs with p less than the maximum p value within this set if we had not conditioned. Note, we list only the most associated SNPs in each LD block by pruning according to LD, as described in Methods.

Disease	Univariate		Conditional		Efficiency ratio
	N.SNPs	FDR	N.SNPs	FDR	
T1D	44	3.79×10^{-6}	4	7.32×10^{-6}	0.11
ATD	3	1.04×10^{-5}	4	1.02×10^{-5}	0.03
CEL	46	3.97×10^{-6}	7	5.77×10^{-6}	0.24
MS	43	5.97×10^{-6}	12	1.19×10^{-5}	0.15
NAR	3	1.61×10^{-4}	0	3.05×10^{-4}	0.83
PBC	25	8.86×10^{-6}	2	2.09×10^{-5}	0.12
PSO	44	2.33×10^{-5}	2	5.45×10^{-5}	0.06
RA	16	2.04×10^{-5}	10	6.00×10^{-5}	0.11
UC	49	4.49×10^{-6}	6	6.73×10^{-6}	0.14
CRO	80	2.08×10^{-6}	12	2.73×10^{-6}	0.28

corresponding study designs in which control groups are independent, and is applicable to a wider range of GWAS datasets for which only summary statistics are available.

Combination of GWAS by analysis of pleiotropy in this sense has several attractive advantages over single-phenotype analysis. The most obvious advantage is improved detection of disease-associated SNPs using GWAS without the need for additional samples. A secondary advantage arises from understanding of the pleiotropic structure between phenotypes: if a SNP is known to exhibit pleiotropy between two conditions, it may be causative for a shared risk factor or pre-disease state. Analysis of such SNPs has the potential to yield information on disease aetiology, with implications for preventative medicine and development of treatment.

A further potential use for this technique could be the genomic analysis of diseases with phenotypes which are difficult to differentiate; for instance, Crohn's disease and Ulcerative Colitis [Farmer et al., 2000]. Additionally, many diseases (such as narcolepsy [Association et al., 1990]) are definitively diagnosed on clinical grounds. This implies that these diseases may constitute a range of biochemical and genetic states. Inclusion criteria based on objective biochemical grounds, such as that used for narcolepsy in the context of this paper [Faraco

Table 2.3 Eight SNP associations with the indicated disease discovered by cFDR but not previously published to our knowledge, together with the phenotype upon which they have been conditioned (C. phen.) and nearby genes. SNPs are shown by RSID: major>minor alleles. The disease (principal phenotype) p value has been corrected for genomic inflation. Note that a SNP reaching significance by $\widehat{cFDR}$ for the principal phenotype does not constitute evidence of association with the conditional phenotype (C. phen.). Chr = chromosome, Pos = position (build GRCh37), SNPs are shown with major>minor alleles, MAF = minor allele frequency in UK controls.

SNP	Chr	Pos	Disease	p value
rs6691768: A>G	1p31.3	61791863	CEL	8.56×10^{-8}
rs1034920: T>C	1p13.1	117076399	PBC	1.43×10^{-6}
rs6705577: G>C	2p21	43359275	CEL	3.61×10^{-7}
rs79248157: T>C	2q32.1	185501065	UC	1.53×10^{-7}
rs72928038: G>A	6q15	90976768	RA	5.89×10^{-7}
rs12924003:C>T	16q12.1	51080214	CRO	5.82×10^{-8}
rs6032606:G>C	20q13.12	44596207	CEL	1.31×10^{-7}
rs9610686:G>A	22q13	37633851	CEL	1.90×10^{-7}

SNP	Chr	Disease	C. phen.	MAF	Nearby Genes
rs6691768: A>G	1p31.3	CEL	CRO	0.373	*NFIA*
rs1034920: T>C	1p13.1	PBC	MS	0.100	*CD58*
rs6705577: G>C	2p21	CEL	MS	0.272	*THADA*
rs79248157: T>C	2q32.1	UC	CRO	0.031	*ZNF804A*
rs72928038: G>A	6q15	RA	T1D	0.176	*BACH2*
rs12924003:C>T	16q12.1	CRO	NAR	0.321	*NOD2*
rs6032606:G>C	20q13.12	CEL	MS	0.052	*ZNF355*, *MMP9*
rs9610686:G>A	22q13	CEL	ATD	0.387	*RAC2*

et al., 2013] are unlikely to characterise all patients with these diseases, and conclusions drawn from studies will not necessarily be medically applicable to the whole patient population. Given this, diseases defined phenotypically with potential genomic diversity may be better analysed by separate consideration of biochemically-defined subtypes, with a collective analysis performed by a method such as $\widehat{cFDR}$, avoiding the assumption that the genomic bases of disease subtypes are identical.

I identify a somewhat counter-intuitive property that the FDR in the union of all regions with $\widehat{cFDR}$ less than a given α tends to be greater than α, and propose a method to overcome this problem. They are complex to apply, but could be much simplified if interest was directed to SNPs with conditional p values less than some threshold p_0, in which my method for allowing for shared controls would ensure that the expected false discovery rate at SNPs

with $\widehat{cFDR}(p_i|p_0) \leq \alpha$ would indeed be controlled at α. My more complicated method to control FDR is necessary if the variable p_j is used in place of the constant p_0. However, I argue in chapter 7, section 7.1.4 that the full $\widehat{cFDR}$ procedure (with variable p_j) is preferable on grounds of efficient information use.

An important consideration in both our method and the original method is that a $\widehat{cFDR}(p_i|p_j)$ value which reaches significance does not constitute genome-wide evidence of association with the conditional phenotype j; indeed, the probability of association with the conditional phenotype relates to $\widehat{cFDR}(p_j|p_i)$ and in general $\widehat{cFDR}(p_i|p_j) \neq \widehat{cFDR}(p_j|p_i)$. In some cases, where the principal p value is very close to genome-wide significance, even conditioning on $p_j \leq 0.5$ can theoretically be enough to reach the relevant $\widehat{cFDR}$ threshold. This is not a weakness of the $\widehat{cFDR}$ method as such, but a consequence of introducing another source of randomness (P_j) to the overall procedure. Principal p values greater than 5×10^{-8} which can be declared significant conditioning on large conditional p value cutoffs correspond to an increase in the area of the region L (see results section), which is accounted for by our FDR-controlling method. This problem is discussed further in chapter 3, section 3.4.

My method enables improved detection of SNPs compared to analysis of unconditional FDR (principal p value alone). However, the improvement is smaller than that reported by Andreasson et al [Andreassen et al., 2013, Andreassen et al., 2014, Andreasson et al., 2014], who detected almost twice as many SNPs using *cFDR* as they would have detected with *uFDR*. This is expected for two reasons. Firstly, the gain in power from *cFDR* essentially comes from an increase in the total number of controls and the effective number of cases. If controls are already shared, the only information gain can come from increasing the number of effective cases. Consequently, the difference in power between *cFDR* and *uFDR* will not be as large when controls are shared, although both outperform their counterparts when controls are split. Secondly, we were careful to use stringent cutoffs for FDR which were chosen to mirror the established genome-wide significance threshold of $p \leq 5 \times 10^{-8}$, generally equivalent to a false discovery rate around 5×10^{-6} to 5×10^{-5}, compared to the more liberal thresholds used by Andreasson et al.

One alternative way to exploit pleiotropic relationships is by meta-analysing two related diseases together, as though the diseases were the same. Our method confers several advantages over this approach. The most important of these is that our method borrows strength from other SNPs according to the level of genome wide pleiotropy between diseases; that is, if the two GWAS suggest extensive pleiotropy (such as Figure 2.4 for T1D | RA), a low p value for a conditional phenotype will 'sway' our judgement of association with the principal phenotype more than the same p value for a conditional phenotype with poor pleiotropy (such

as Figure 2.4, for T1D | PSO). A meta-analysis would not distinguish these two scenarios. A secondary advantage of my technique is that SNP detection is not systematically weakened if the two diseases do not exhibit pleiotropy, as would be the case in meta-analysis; this arises because I am testing association with only one of the two phenotypes at a time.

2.4 Methods

2.4.1 Datasets

I obtained SNP summary statistics from ten studies on autoimmune diseases from ImmunoBase (`www.immunobase.org`). Inclusion and exclusion criteria for the studies are described in detail in the original publications ([Onengut-Gumuscu et al., 2014, Cooper et al., 2012, Trynka et al., 2012, Beecham et al., 2013, Faraco et al., 2013, Liu et al., 2012, Tsoi et al., 2012, Eyre et al., 2012, Jostins et al., 2012, Jostins et al., 2012]. Generally, some or all controls from different studies were obtained from common data sources, resulting in overlapping control groups. All studies used the ImmunoChip array [Cortes and Brown, 2011].

P values for type 1 diabetes were from a meta-analysis of a case-control study and familial study using the transmission disequilibrium test (TDT). In order to calculate the correlation between p values for different diseases, it was necessary to calculate effective numbers of cases and controls for the combined T1D study. For a case control study, under the assumptions of Hardy-Weinberg and the null hypothesis, the variance of the log odds ratio may be expressed as

$$\frac{n_0+n_1}{2n_0n_1}\frac{1}{f(1-f)}$$

where n_0 and n_1 are the numbers of cases and controls and f is the minor allele frequency in controls.

Given the standard error of a log OR for the TDT study, $\hat{\sigma}$, and a minor allele frequency, we estimated $M = \hat{\sigma}^2 f(1-f)$ for all ImmunoChip SNPs which did not show deviation from the null hypothesis ($p > 0.5$). The distribution of $\log(M)$ is shown in appendix A, figure A.7. By equating the median of M with $\frac{n_0+n_1}{n_0n_1}$, and assuming that each TDT family contributed the equivalent information to one control in a case-control study, ie $n_0 = 2943$, an equivalent number of cases was estimated to be 4126. This seemed reasonable, given that there were a total of 5505 (dependent) cases across those families.

SNPs were excluded on the basis of QC summaries calculated on 12,888 common controls: call rate less than 99%, minor allele frequency less than 0.02, or deviation from Hardy-Weinberg equilibrium ($|Z| > 5$). Given the strong association of immune mediated diseases with the MHC and the extended LD in the region, we were concerned that MHC SNPs might cause inaccurate estimation of pleiotropy. I therefore excluded SNPs in a wide band around the MHC region on chromosome 6 (co-ordinates 24500000: 34800000, build NCBI36). After quality control, genotype data was available for at least one phenotype at a total of 110677 SNPs.

2.4.2 Genomic control

P values were corrected for genomic inflation using a genomic control algorithm [Devlin et al., 2001]. A set of SNPs known to be un-associated with autoimmune disease was obtained from the Wellcome Trust Case Control Consortium (WTCCC) study on reading and mathematics ability. These SNPs were pruned so that none were in LD with $r^2 > 0.2$, and any SNPs within 500 kb of known autoimmune-associated regions were removed. The average degree of inflation was computed for each disease at the remaining 1761 SNPs, and all effect sizes and p values were adjusted accordingly.

2.4.3 Computation of expected quantile

From the first part of (2.2), we have:

$$cFDR(p_i|p_j) = \frac{Pr(H_0^{(i)}|P_j \leq p_j)Pr(P_i \leq p_i|P_j \leq p_j, H_0^{(i)})}{Pr(P_i \leq p_i|P_j \leq p_j)} \tag{2.10}$$

As per Andreasson et al [Andreassen et al., 2013], the quantity $Pr(H_0^{(i)}|P_j \leq p_j)$ is set conservatively at 1, and the quantity $Pr(P_i \leq p_i|P_j \leq p_j)$ is estimated empirically as the proportion of pairs of observed p values (p'_i, p'_j) with $p'_j \leq p_j$ which also satisfy $p'_i \leq p_i$.

For a given SNP, let η denote the standardised mean allele frequency (MAF) difference; that is, the Z value we would compute if the observed MAFs agreed exactly with the population MAFs. I consider η for a random SNP as being an instance of a random variable H, and that the observed z value for that SNP $Z|H = \eta$ is distributed as

$$Z|H = \eta \sim N(\eta, 1) \tag{2.11}$$

We further assume that H follows a mixture distribution taking the value 0 with probability $\pi_0^{(j)}$ and a normal pdf with probability $1-\pi_0^{(j)}$:

$$H \sim \begin{cases} 0, & p = \pi_0^{(j)} \\ N(0,\sigma^2), & p = 1-\pi_0^{(j)} \end{cases} \tag{2.12}$$

This implies

$$Z_j \sim \begin{cases} N(0,1), & p = \pi_0^{(j)} \\ N(0,1+\sigma^2), & p = 1-\pi_0^{(j)} \end{cases} \tag{2.13}$$

Thus, given the observed distribution of Z_j, the parameters $\pi_0^{(j)}$ and σ_j may be estimated by an expectation - maximisation algorithm (https://gist.github.com/chr1swallace/11421212).

We assume as per Zaykin [Zaykin and Kozbur, 2010] that the distribution of pairs of observed z values (Z_i, Z_j) for a single given SNP is bivariate normal. Denote by $H_\eta^{(j)}$ the event that, for a given SNP, the values Z_j are distributed as $N(\eta,1)$, with η depending on the SNP.

Under our assumption of the null hypothesis $H_0^{(i)}$ for the principal phenotype and a population MAF difference corresponding to η for the conditional phenotype, we have

$$(Z_i, Z_j | H_0^{(i)}, H_\eta^{(j)}) \sim N\left(\begin{pmatrix} 0 \\ \eta \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{2.14}$$

The correlation ρ arises from the shared controls between groups [Zaykin and Kozbur, 2010, Lin and Sullivan, 2009] and is asymptotically equal to

$$\rho = \frac{1}{\sqrt{\left(1+\frac{N_{0i}}{N_0}\right)\left(1+\frac{N_{0j}}{N_0}\right)\left(1+\frac{N_{0i}}{N_i}+\frac{N_0}{N_i}\right)\left(1+\frac{N_{0j}}{N_j}+\frac{N_0}{N_j}\right)}} \tag{2.15}$$

where N_i and N_j are the numbers of cases, N_{0i} and N_{0j} are the numbers of non-shared controls, and N_0 is the number of shared controls for the original GWAS for the principal and conditional phenotypes respectively. There is good agreement with the asymptotic correlation when group sizes are greater than 100 [Zaykin and Kozbur, 2010].

Given equations (2.12)–(2.15), the joint distribution of Z_i and Z_j can be computed under only the assumption $H_0^{(i)}$. The value of the partial PDF of $(Z_i, Z_j | H_0^{(i)})$ at (x,y) can be derived in a similar way to (2.13):

$$\left(Z_i, Z_j | H_0^{(i)}\right) \sim \begin{cases} N\left(\binom{0}{0}, \binom{1\ \rho}{\rho\ 1}\right), & p = \pi_0^{(j)} \\ N\left(\binom{0}{0}, \binom{1\ \ \rho}{\rho\ 1+\sigma^2}\right), & p = 1 - \pi_0^{(j)} \end{cases} \tag{2.16}$$

We now compute the final probability in equation (2.10). Define

$$P_\eta(X) = Pr(X | H_0^{(i)}, H_\eta^{(j)}) \tag{2.17}$$

as the probability of observing events X for a particular SNP with true effect size η (which may be 0, corresponding to the general null). Then,

$$Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) = \frac{Pr(P_i \leq p_i, P_j \leq p_j | H_0^{(i)})}{Pr(P_j \leq p_j | H_0^{(i)})}$$

$$= \frac{\pi_0^{(j)} P_0(P_i \leq p_i, P_j \leq p_j) + (1 - \pi_0^{(j)}) \int_{-\infty}^{\infty} P_\eta(P_i \leq p_i, P_j \leq p_j) f(\eta) d\eta}{\pi_0^{(j)} P_0(P_j \leq p_j) + (1 - \pi_0^{(j)}) \int_{-\infty}^{\infty} P_\eta(P_j \leq p_j) f(\eta) d\eta}. \tag{2.18}$$

If the distribution of H is estimable by other means, quantity (2.18) can be calculated numerically without the assumption that the non-null component of H be normally distributed, at the cost of higher computation time. Under our assumptions, equations (2.13) and (2.16) enable the fast computation of quantity (2.18) by normal CDFs; writing

$$\begin{aligned} \Lambda_{(\rho,\sigma^2)}(z_i, z_j) &= \int_{|x|>|z_i|, |y|>|z_j|} N_{\binom{0}{0}, \binom{1\ \ \rho}{\rho\ 1+\sigma^2}}(x, y) dx dy \\ \lambda_{\sigma^2}(z_j) &= \int_{|y|>|z_j|} N_{(0,1+\sigma^2)}(y) dy \end{aligned} \tag{2.19}$$

we have

$$Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) = \frac{\pi_0^{(j)} \Lambda_{(\rho,0)}(z_i, z_j) + (1 - \pi_0^{(j)}) \Lambda_{(\rho,\sigma^2)}(z_i, z_j)}{\pi_0^{(j)} \lambda_0(z_j) + (1 - \pi_0^{(j)}) \lambda_{\sigma^2}(z_j)} \tag{2.20}$$

2.4.4 Point expected quantile

Because the formula for $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ is differentiable on the unit square, an expression for the expected quantile of p_i given an exact value for p_j can be computed by taking the partial derivative with respect to p_j:

$$Pr(P_i \leq p_i | P_j = p_j, H_0^{(i)}) = \frac{\partial}{\partial p_j} Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) = \frac{A+B}{C}$$

where

$$A = \pi_0^{(j)} N_{(0,1)}(z_j) \int_{|x| \geq z_i} N_{(\rho z_j, 1-\rho^2)}(x) dx$$
$$B = (1-\pi_0^{(j)}) N_{(0,1+\sigma^2)}(z_j) \int_{|x| \geq z_i} N_{\left(\frac{\rho z_j}{1+\sigma^2}, \frac{1-\rho^2+\sigma^2}{1+\rho^2}\right)}(x) dx$$
$$C = \pi_0^{(j)} N_{(0,1)}(z_j) + (1-\pi_0^{(j)}) N_{(0,1+\sigma^2)}(z_j)$$

where $N_{\mu,\sigma^2}(x)$ denotes the value of the normal pdf with mean μ and variance σ^2 at x.

2.4.5 Significance thresholds

Because $\widehat{uFDR}$ values are monotonically related to p values, the widely accepted GWAS p value cutoff of 5×10^{-8} corresponds naturally to a cutoff for $\widehat{uFDR}$. For each phenotype i, I set a significance threshold β^i for $\widehat{uFDR}(p_i)$ as the lowest possible value of γ for which $\widehat{uFDR}(p_i) \leq \gamma \Leftrightarrow p_i \leq 5 \times 10^{-8}$.

I then applied an analogous approach to $\widehat{cFDR}$. For each pair of phenotypes (i, j), I set a significance threshold α_j^i as the value of γ' for which $\widehat{cFDR}(p_i|p_j) \leq \gamma' \Leftrightarrow p_i \leq 5 \times 10^{-8}$. Given the distribution of P_j, it is possible that this could lead to declaring SNPs with $p_i > 5 \times 10^{-8}, p_j \approx 1$ as significant. To avoid this, if α_j^i was larger (less stringent) than β^i, I set $\alpha_j^i = \beta^i$.

For each ordered pair of phenotypes (i, j), I declared all SNPs with $\widehat{cFDR}(p_i|p_j) \leq \alpha_j^i$ as non-null for phenotype i. This included all SNPs with $\widehat{uFDR}(p_i) \leq \beta^i$. I then computed bounds $c_j^{(i)}$ on the false discovery rate amongst SNPs for which $\widehat{cFDR}(p_i, p_j) \leq \alpha_j^i$ as per section 2.2.5. For each phenotype, this gave nine upper bounds, corresponding to each of the nine conditional phenotypes.

2.4.6 Network and heatmap representation of pleiotropy

I compared the degree of pleiotropy between diseases by considering how much the p-value threshold for significance for the principal phenotype changed when conditioning on a small p-value threshold for the conditional phenotype. I used the $\widehat{cFDR}$ algorithm to compute the number p_i^{j*} such that $P(H_0^{(i)} | P_i \leq p_i^{j*}, p_j \leq 5 \times 10^{-6}) = P(H_0^{(i)} | P_i \leq 5 \times 10^{-8})$; that is,

$\widehat{cFDR}(p_i^{j*}|5\times10^{-6}) = \widehat{uFDR}(5\times10^{-8})$. I then considered the ratio $p_i^{j*}/5\times10^{-8}$; that is, the fold increase in significance cutoff after conditioning.

I visualised the ratio $p_i^{j*}/5\times10^{-8}$ as a heat-map (appendix A, figure A.6). I also produced a network (Figure 2.6), with an edge from vertex i to vertex j if and only if, by conditioning on $P_j \leq 5\times10^{-6}$, the cutoff for significance for P_i could be increased from 5×10^{-8} to 4×10^{-7}. This cutoff was chosen as the smallest value such that the network was weakly connected; that is, each vertex had an arrow either to it or from it.

2.4.7 Discovery of novel SNP associations

SNPs were deemed significant for each principal phenotype i if $\widehat{cFDR}(p_i|p_j) \leq \alpha_j^i$ for any conditional phenotype j and $\alpha_j^i \leq \beta^i$. The list of SNPs declared non-null for phenotype i was pruned to allow for linkage disequilibrium (LD) by listing all SNPs in increasing order of $\min_j(\widehat{cFDR}(p_i|p_j)$ and stepping through the list from left to right, at each stage removing all SNPs in LD with $r^2 \geq 0.1$ to the right of the current SNP. This ideally leads to the inclusion of at most one SNP from each LD block.

2.4.8 Multiple Testing

A multiple testing problem arises from considering p values for one disease conditioned separately on nine others. Specifically, if the criterion for declaring a SNP non-null for phenotype i is that $\widehat{cFDR}(p_i|p_j) \leq \alpha_j^i$ for at least one of the nine possible values of j, then the FDR for all SNPs declared non-null will be greater than the FDR among the smaller set of SNPs for which $\widehat{cFDR}(p_i|p_j) \leq \alpha_j^i$ for only one value of j, due to multiple testing.

However, this excess FDR is not enough to warrant a Bonferroni (Sidak) correction; the $\widehat{cFDR}(p_i|p_j)$ values for a phenotype i are highly correlated, as all are in turn highly correlated with p_i. A Bonferroni correction tends to remove any advantage in SNP detection gained from $\widehat{cFDR}$, though an advantage may still be seen when only considering one conditional phenotype j.

A method proposed by Nyholt [Nyholt, 2004] was used to correct for multiple testing in SNPs with high LD. I estimated a correlation matrix Ω for potentially non-null cFDR values using Spearman's rank correlation. The variance of the eigenvalues of Ω, $\mathrm{var}(\lambda_{obs})$, was computed and used to estimate the effective number of variables M_{eff} according to the equation

$$M_{eff} = 1 + 9\left(1 - \frac{\mathrm{var}(\lambda_{obs})}{9}\right) \tag{2.21}$$

Note that var(λ_{obs}) is at most 9 (completely correlated variables, effectively one test) and at least 0 (completely uncorrelated variables, essentially a Bonferroni correction).

Denoting by n_j^i the number of SNPs with $\widehat{cFDR}(p_i|p_j) \leq \alpha_j^i$, corresponding to an upper bound on the FDR of c_j^i, an upper bound for the FDR among all SNPs declared significant for phenotype i was then computed as

$$c_0^i = M_{eff} \frac{\sum_{j=1..10, j\neq i} c_j^i n_j^i}{\sum_{j=1..10, j\neq i} n_j^i}, \tag{2.22}$$

intuitively, multiplying the expected average number of false discoveries across conditional phenotypes ($c_j^i n_j^i$) by the effective number of tests. Values of M_{eff} and c_0^i are shown in appendix A, table A.2.

Chapter 3

Applications of cFDR method

3.1 Introduction

The cFDR method and the extensions derived in the previous chapter are widely applicable, requiring only genome-wide summary statistics. However, the method is still relatively new, particularly the application to shared control designs and control of overall FDR introduced in chapter 2. This chapter details two additional applications of cFDR on new disease datasets.

Firstly, I reapplied the method on a previously published genetic dataset for juvenile idiopathic arthritis (JIA) [Hinks et al., 2013] , conditioning on the clinically related diseases RA and T1D. Application of the cFDR method enabled improved genetic discovery while maintaining FDR control. Section 3.2.1 gives an overview of the JIA phenotype, and avenues for investigation which could lead to improved treatment strategies. Genetic analysis of JIA is discussed further in chapter 6.

Secondly, I applied the cFDR method to a new GWAS dataset on the phenotype of Eosinophilic Granulomatosis with Polyangiitis (EGPA), levering on large GWAS for eosinophil count [Astle et al., 2016] (EC) and asthma [Moffatt et al., 2010], and a smaller concurrent GWAS on anti-neutrophil cytoplasmic antibody-associated vasculitis (AAV). This was able to substantially improve discovery of SNP-disease associations, thus improving understanding of the pathophysiology of EGPA and related conditions.

These two analyses represent different areas of application for the cFDR method. In the analysis of JIA in section 3.2, the traits for leverage (RA and T1D) were chosen due to known comorbidity and several known similarities in genetic basis, although they are notably distinct diseases from JIA itself[1]. In EGPA (section 3.3), I was able to lever on traits

[1]The nomination of 'similar' traits to study is obviously subjective, but the systematic selection of levering traits is beyond the scope of this thesis

which were more related to the disease; in effect, constituents of the disease itself (strictly, the diagnosis of EGPA requires a pathologically raised eosinophil count, and EC considers only normal eosinophil counts). The traits used for leverage also differed in their type; EC is a haematological measurement in healthy individuals rather than a disease trait such as T1D and RA.

Both applications were on GWAS in diseases of relatively lower prevalence than typical GWAS traits, and hence had limited sample size, incentivising the use of methods such as cFDR. As well as improving the potential for SNP discovery, these applications gave further insight into the practicalities of applying the cFDR method, particularly regarding the choice of threshold for association. The analysis of EGPA illustrated the potential for using cFDR to study rare phenotypes by levering on very large datasets for common related traits.

3.2 Applications to juvenile idiopathic arthritis

3.2.1 JIA and traits for conditioning

JIA is an autoimmune disorder of childhood (onset at age < 16) affecting around one in 1000 children per year. It is diagnosed symptomatically based on the onset of non-infective joint inflammation persisting for six weeks without an established alternative cause. Patients frequently present with symptoms typical of inflammation, including fever and lethargy. Extra-articular autoimmune processes may also be present, including onycholysis, dactylitis, and iridocyclitis [Ravelli and Martini, 2007]. Collectively, these characteristics of the disease and its symptomatic diagnosis suggest that a range of inflammatory and other processes are implicated in JIA pathogenesis. The disease is sub-classified into a range of subtypes, discussed in chapter 6, section 6.1. The differential pathology of JIA subtypes is an important topic explored in chapter 6, but within-disease genetic heterogeneity is not considered in this chapter.

As a degenerative disease of childhood, JIA has a significant disease burden on patients and on the healthcare system. The disease can be painful and debilitating, and greatly reduce quality of life. Children with JIA frequently have retarded growth and failure to thrive, due to both articular damage and systemic inflammation [Ravelli and Martini, 2007], and are more likely to develop other inflammatory conditions including uveitis and coronary vascular disease [Raab et al., 2012]. Children with JIA are restricted in their ability to undertake activities of daily living, frequently into adulthood, placing considerable stress on parents and other carers [Packham and Hall, 2002].

Despite a recently improved recent range of treatment options, there is potential for considerable development in the evidence-based management of JIA, both in the development of new pharmacological therapies and the improved determination of which treatments are likely to be effective in a given patient [Ravelli and Martini, 2007]. Timely effective therapy is of particular importance in JIA given the long-term impact of the disease on the patient's joint function and development. Understanding risk factors for JIA and identifying patients at risk of the disease also has therapeutic benefit. A promising line of investigation towards these aims is in studying the genetic architecture of the disease.

Genetics of JIA

Determination of genetic associations of JIA has been slower than for phenotypes with comparable childhood incidence such as type 1 diabetes, partly because of difficulties in recruitment due to the relatively recent diagnostic criteria and the comparative heterogeneity of the phenotype, and partly because of lower prevalence (0.1-0.2% [Duurland and Wedderburn, 2014] compared to around 0.6% in T1D [Hex et al., 2012]). A GWAS published in 2012 on 814 cases was the largest at that time, and was restricted to oligoarticular and RF-negative polyarticular subtypes in order to generate homogeneous study cohorts [Thompson et al., 2012]. No loci reached genome-wide significance. A range of candidate-gene studies associated the *PTPN2*, *PTPN22*, and MHC regions with JIA [Hinks et al., 2013]

The genomic investigation of JIA was continued with an ImmunoChip study [Hinks et al., 2013] on 2,816 cases and 13,056 controls. Although the ImmunoChip was not explicitly designed to cover putative JIA-associated regions, the fine-mapping of general autoimmunity-associated loci was expected to cover several undiscovered JIA regions. Like the earlier GWAS, the ImmunoChip study was restricted to the oligoarthritis and RF-negative polyarthritis subtypes. The study found 14 additional JIA-associated loci at genome-wide significance.

JIA is known to have comorbidity with T1D [Hermann et al., 2015], with which the *PTPN22*, *PTPN2* and MHC regions are also known to be associated [Bottini et al., 2004, Todd et al., 1987, Barrett et al., 2009]. A non-parametric comparison of p-value rankings [Burren et al., 2014] found that, based on ImmunoChip data, JIA was more similar to T1D than 14 other autoimmune conditions [Onengut-Gumuscu et al., 2014]. This suggested that leverage on T1D may improve genetic discovery. The same could be expected of RA given the phenotypic similarity of the disease to JIA. Given this potential, I used the data from the JIA ImmunoChip study to conduct an analysis using cFDR, conditioning on data from ImmunoChip studies on T1D [Onengut-Gumuscu et al., 2014] and RA [Eyre et al., 2012].

My aim was to examine whether additional JIA-associated variants could be identified by levering these three related diseases.

3.2.2 Methods

Quality control

SNPs were included in each pairwise analysis (JIA|T1D, JIA|RA) if they passed quality control criteria in both of the original studies (of [Onengut-Gumuscu et al., 2014, Eyre et al., 2012, Hinks et al., 2013]). In addition, SNPs were excluded from the analysis if their MAF across cases and controls was $\leq 2\%$, if their overall call rate in either cases or controls was $\leq 99\%$, or if they deviated from Hardy-Weinberg equilibrium across cases and controls with p-value $\leq 2\Phi(-4) = 6.3 \times 10^{-5}$ in either of the two datasets used in each analysis. The MHC region was removed from the analysis with wide margins (24.5-34.8 Mb, chromosome 6; NCBI build 37). After quality control, 111,406 SNPs were available for the JIA/T1D analysis and 100,611 for the JIA/RA analysis.

Observed summary statistics were substantially inflated (JIA: $\lambda = 1.67$, T1D: $\lambda = 1.64$, RA: $\lambda = 1.54$); however, due to the restricted cover of the ImmunoChip, these values were likely to be overestimates of inflation due to a high frequency of genuine associations. Following earlier work with the ImmunoChip [Trynka et al., 2012, Onengut-Gumuscu et al., 2014], inflation factors were computed using a set of SNPs expected to have no effect for autoimmune phenotypes (see chapter 2, section 2.2.6). Using this restricted set of SNPs, the inflation statistics were more reasonable (JIA: $\lambda = 1.16$; T1D: 1.11; RA: 0.98). The dataset was adjusted for this residual inflation by scaling χ^2 test statistics [Devlin et al., 2001].

Adjustment for shared controls

The presence of shared controls between studies was adjusted for using the procedure described in chapter 2. Total sample sizes and estimated parameters of Z-score distributions were as follows:

Table 3.1 Study sizes and parameters of effect size distributions for JIA, T1D and RA ImmunoChip studies

	Controls	Cases	Est. π_0	Est. σ
JIA	13056	2816	0.68	1.89
T1D	12228	6670	0.78	2.98
RA	15870	11475	0.78	5.77

I estimated an overlap of 8530 controls in both analyses, leading to correlations of 0.17 and 0.16 between Z-scores for null SNPs in the JIA|T1D and JIA|RA analyses respectively.

Conditioning on type 1 diabetes and rheumatoid arthritis

I drew Q-Q plots showing observed and expected distributions of summary statistics for each of the four analysis (JIA|T1D, T1D|JIA, JIA|RA, RA|JIA) using an adaptation of the procedure described in [Andreassen et al., 2013] and used in chapter 2, section 2.2.6. As in the previous chapter, I chose to show observed and expected quantiles of 'adjusted' p-values p'_{JIA}, p'_{T1D}, p'_{RA} rather than raw p-values p_{JIA}, p_{T1D}, p_{RA} (using the notation and method from chapter 2, sections 2.2.3, 2.4.3):

$$p'_{JIA} = Pr(P_{JIA} < p_{JIA} | P_{T1D} < p_{T1D}, H_0^{JIA}) \ / \ Pr(P_{JIA} < p_{JIA} | P_{RA} < p_{RA}, H_0^{JIA})$$
$$p'_{T1D} = Pr(P_{T1D} < p_{T1D} | P_{JIA} < p_{JIA}, H_0^{T1D}) \ / \ Pr(P_{T1D} < p_{T1D} | P_{RA} < p_{RA}, H_0^{T1D})$$
$$p'_{RA} = Pr(P_{RA} < p_{RA} | P_{JIA} < p_{JIA}, H_0^{RA}) \ / \ Pr(P_{RA} < p_{RA} | P_{T1D} < p_{T1D}, H_0^{RA})$$

noting that the definitions of $p'_{JIA}, ...$ differ between analyses. P-values for the conditional phenotype were unadjusted. Although in practice making only a very subtle difference to the appearance of the plots, this corrects the Q-Q plots for shared controls between studies. Q-Q plots showed good evidence for increasing inflation with conditioning for all four analyses (figures 3.1).

Two thresholds for cFDR significance were considered, each of which corresponded to the standard genome-wide p-value significance threshold of 5×10^{-8} in different ways. Denoting by p_i^A, p_i^B the p-values for SNP i for traits A, B respectively, the cutoff $\alpha_1^{A|B}$ on the cFDR values in the analysis $A|B$, $\widehat{cFDR}(p^A|p^B)$, was set as

$$\alpha_1^{A|B} = \max_{i | p_i^A \leq 5 \times 10^{-8}} \left(\widehat{cFDR}(p_i^A | p_i^B) \right) \tag{3.1}$$

The FDR of the procedure in which SNPs are declared non-null by $\widehat{cFDR}(p^A|p^B) \leq \alpha_1^{A|B}$ is generally higher than the FDR of the procedure in which SNPs are declared non-null by $p_i \leq 5 \times 10^{-8}$ (and larger than $\alpha_1^{A|B}$ itself) so the second, more conservative threshold $\alpha_2^{A|B}$ was declared such that upper bounds on the FDR values

$$Pr\left(H_0^A | \widehat{cFDR}(p^A|p^B) \leq \alpha_2^{A|B} \right) \tag{3.2}$$
$$Pr\left(H_0^A | p_i^A \leq 5 \times 10^{-8} \right) \tag{3.3}$$

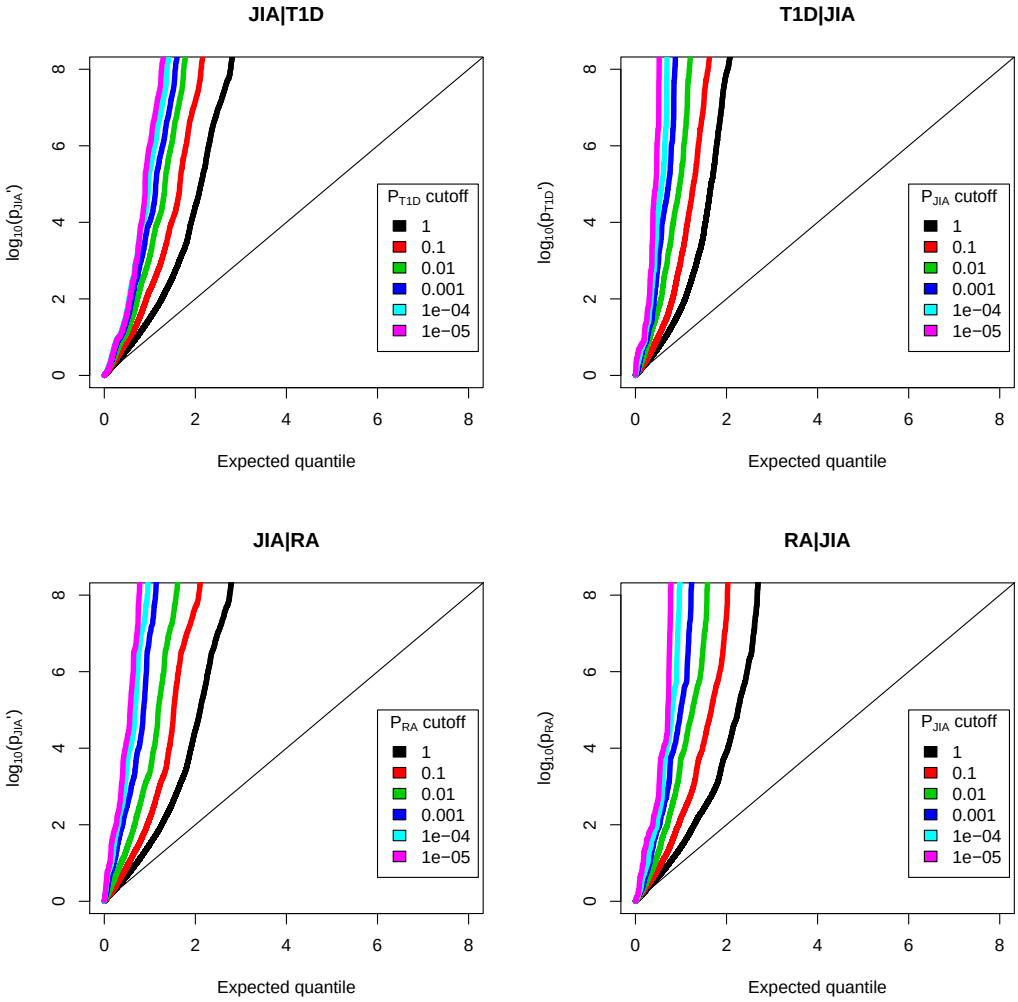


Fig. 3.1 Conditional Q-Q plots showing inflation of summary statistics for one phenotype conditioning on another. If there are a body of SNPs which show pleiotropy between the principal and conditional phenotypes, the deviation of the Q-Q plot leftward of the X-Y line should increase with conditioning on successively lower p-value thresholds for the conditional phenotype. Inflation of summary statistics is evident in all cases when conditioning on p-value thresholds in the conditional phenotype.

were equal. An upper bound on quantity 3.2 was estimated using the procedure described in chapter 2, section 2.2.5, and an upper bound on quantity 3.2.2 using the Benjamini-Hochberg procedure. Thresholds are given in table 3.2 below.

Table 3.2 Thresholds on $\widehat{cFDR}$ for analysis of JIA, T1D and RA

	$\alpha_1^{A\|B}$	FDR bound	$\alpha_2^{A\|B}$	FDR bound
JIA|T1D	1.00×10^{-5}	5.59×10^{-5}	2.97×10^{-6}	1.79×10^{-5}
JIA|RA	1.36×10^{-5}	7.28×10^{-5}	2.55×10^{-6}	1.71×10^{-6}
T1D|JIA	3.96×10^{-6}	9.04×10^{-6}	1.95×10^{-6}	4.94×10^{-6}
RA|JIA	1.26×10^{-5}	5.47×10^{-5}	4.48×10^{-6}	2.06×10^{-5}

SNPs with $\widehat{cFDR}$ values less than the threshold were pruned to remove SNPs in LD with $r^2 \geq 0.1$ using the protocol described in chapter 2, section 2.4.7, prioritising SNPs by $\widehat{cFDR}$ value.

3.2.3 Results

The analysis of JIA conditioning on T1D yielded 12 non-MHC variants for which $\widehat{cFDR} \leq \alpha_1^{JIA|T1D}$ and 8 for which $\widehat{cFDR} \leq \alpha_2^{JIA|T1D}$. Five variants with $\widehat{cFDR} \leq \alpha_1^{JIA|T1D}$ and 1 with $\widehat{cFDR} \leq \alpha_2^{JIA|T1D}$ were not genome-wide significant by p-value alone ($P(JIA) > 5 \times 10^{-8}$). Results are shown in table 3.3.

Table 3.3 Results for cFDR analysis of JIA|T1D. Only results for SNPs not in LD with a SNP reaching GWS-by-p-value in JIA by p-value are shown. * - $\widehat{cFDR} \leq \alpha_2^{JIA|T1D}$

rsID	Chr.	Pos.	P(JIA)	P(T1D)	$\widehat{cFDR}$
rs7808122	7	22764605	5.8×10^{-08}	0.0551	6.35×10^{-06}
rs34132030	13	41954036	1.77×10^{-07}	0.0164	9.78×10^{-06}
rs79893749*	3	46228654	1.88×10^{-07}	1.54×10^{-05}	2.58×10^{-06}
rs66718203	16	11336144	4.46×10^{-07}	2.36×10^{-07}	3.68×10^{-06}
rs6740838	2	100179931	8.83×10^{-07}	3.29×10^{-06}	8.31×10^{-06}

SNP rs79893749 is near the genes encoding CCR1 and CCR3, which are chemokine receptors and are good candidates for JIA association. SNP rs66718203 (16p13.13) is near the *CLEC16A* region, known to be associated with T1D, MS, CEL, PSO, IBD and PBC. SNP rs6740838 is near the AFF3 gene, known to be associated with RA, T1D, and CEL.

The analysis for JIA|RA yielded six SNPs not in LD with SNPs reaching GWS-by-p-value for JIA, shown in table 3.4. SNP rs7993214, in the *COG*6 region, is in LD ($r^2 > 0.2$) with the

imputed SNP rs9532434, which reached genome-wide significance for JIA ($p = 4.52 \times 10^{-8}$). SNP rs79893749 is in a region associated with CEL and T1D, containing the $CCR1/CCR3$ genes. SNP rs6740838 is in a region associated with RA, T1D and CEL, containing the $AFF3$ gene. SNP rs2364480 is in a region containing an imputed SNP (rs10849448) reaching genome-wide significance for JIA ($p = 4.54 \times 10^{-9}$). SNP rs5029924 is in a region associated with multiple autoimmune conditions containing the genes $OLIG3$ and $TNFAIP3$. Finally, SNP rs4755450 is in a region containing the imputed SNP rs7127214 reaching genome-wide significance ($p = 1.90 \times 10^{-8}$).

Table 3.4 Results for cFDR analysis of JIA|RA. Only results for SNPs not in LD with a SNP reaching GWS-by-p-value in JIA by p-value are shown. * - $\widehat{cFDR} \leq \alpha_2^{JIA|RA}$

rsID	Chr.	Pos.	P(JIA)	P(RA)	$\widehat{cFDR}$
rs7993214*	13	39248912	1.61×10^{-07}	1.02×10^{-4}	1.18×10^{-06}
rs79893749	3	46228654	1.88×10^{-07}	1.92×10^{-3}	2.86×10^{-06}
rs6740838	2	100179931	8.83×10^{-07}	5.48×10^{-06}	4.5×10^{-06}
rs2364480	12	6365536	5.1×10^{-08}	7.14×10^{-2}	4.98×10^{-06}
rs5029924	6	138229191	2.86×10^{-06}	3.52×10^{-07}	8.71×10^{-06}
rs4755450	11	36320151	3.35×10^{-07}	2.16×10^{-2}	1.16×10^{-05}

Results for the T1D|JIA analysis yielded three SNPs with $\widehat{cFDR} \leq \alpha_1^{T1D|JIA}$ not in LD with SNPs reaching GWS-by-p-value for T1D, two with $\widehat{cFDR} \leq \alpha_2^{T1D|JIA}$. All three SNPs were in regions which reached genome-wide significance in the published paper, which included additional TDT families [Onengut-Gumuscu et al., 2014] although none did in the (reduced) case-control dataset used in this analysis.

Table 3.5 Results for cFDR analysis of T1D|JIA. Only results for SNPs not in LD with a SNP reaching GWS-by-p-value in T1D by p-value are shown. * - $\widehat{cFDR} \leq \alpha_2^{T1D|JIA}$

rsID	Chr.	Pos.	P(T1D)	P(JIA)	$\widehat{cFDR}$
rs72727394	15	36634314	5.22×10^{-08}	0.0163	1.34×10^{-06}
rs12150079	17	35278943	1.88×10^{-07}	0.0118	3.74×10^{-06}
rs113010081	3	46432416	3.43×10^{-07}	8.11×10^{-07}	1.23×10^{-06}

In the RA|JIA analysis, eight SNPs not GWS-by-p-value reached significance by $\widehat{cFDR} < \alpha_1^{RA|JIA}$, three of which satisfied $\widehat{cFDR} < \alpha_2^{RA|JIA}$. Results are shown in table 3.6. All SNPs in the RA|JIA analysis are in regions found to be RA-associated in a more recent meta-analysis [Okada et al., 2014].

Table 3.6 Results for cFDR analysis of RA|JIA. Only results for SNPs not in LD with a SNP reaching GWS-by-p-value in RA by p-value are shown. * - $\widehat{cFDR} \leq \alpha_2^{RA|JIA}$

rsID	Chr.	Pos.	P(RA)	P(JIA)	$\widehat{cFDR}$
rs3087243*	2	204447164	1.36×10^{-07}	2.68×10^{-04}	1.81×10^{-06}
rs35677470*	3	58158676	1.74×10^{-07}	2.48×10^{-05}	1.58×10^{-06}
rs9979383*	21	35637631	5.23×10^{-07}	1.06×10^{-08}	1.58×10^{-06}
rs595158	11	60666157	1.79×10^{-07}	4.9×10^{-03}	5.87×10^{-06}
rs12936409	17	35297175	3.72×10^{-07}	6.76×10^{-04}	6.00×10^{-06}
rs28532547	1	2551146	2.28×10^{-07}	4.72×10^{-03}	7.23×10^{-06}
rs75351767	7	37393876	2.94×10^{-07}	4.36×10^{-03}	8.87×10^{-06}
rs39984	5	102625191	9.29×10^{-08}	1.06×10^{-01}	1.25×10^{-05}

3.3 Applications to EGPA

3.3.1 EGPA and traits for conditioning

EGPA, formerly called Churg-Strauss syndrome, is characterised by eosinophilic tissue infiltration. The disease typically has a clinical prodrome of asthmatic symptoms, with more serious symptoms developing months to years later. It is a rare condition, with an incidence in Europe of approximately 0.5-0.8 per million per year and a prevalence of 10.7–13 cases per million [Gioffredi et al., 2014].

EGPA is typically considered a small-vessel autoimmune vasculitis. However, many patients have no evidence of vascular involvement and do not have raised levels of anti-neutrophil cytoplasmic antibodies (a marker of systemic vasculitis). This suggests a distinction between vasculitic EGPA, characterised by conditions such as glomerulonephritis and purpura due to small-vessel damage, and eosinophilic EGPA, in which tissue damage such as pulmonary infiltration and cardiomyopathy arise due to tissue eosinophilia.

The recommended treatment for initial occurrence of the disease is typically glucocorticoids, or dual therapy with cyclophosphamide in serious cases. Remission is typically managed with methotrexate or azathioprine [Bosch et al., 2007]. All first-line drugs have marked systemic toxicity, and there is considerable incentive for development of new therapies. This could be facilitated by better understanding of disease pathology through investigation of its genetic basis. In light of the heterogeneous nature of the disease, there may also be scope for precision medicine approaches targeting individuals with particular subtypes.

Eosinophils and eosinophilic disease

Eosinophils are immune cells involved in defence against multicellular parasites, and are involved in the mechanisms of atopy. Typically, eosinophils make up 1-5% of circulating leukocytes, corresponding to a plasma concentration of $< 0.4 \times 10^9 L^{-1}$ in adults. Eosinophilia is a state of raised circulating eosinophil count, typically $> 0.45 \times 10^9 L^{-1}$. This may be idiopathic, secondary to another condition, or in response to outside insult such as an allergen [Simon and Simon, 2007]. Eosinophilia is associated with several clinical syndromes depending on the site of eosinophilic infiltration.

Eosinophil-modulated syndromes can occur even when eosinophils are within the normal range [Straumann et al., 2012]. This suggests that amongst asymptomatic patients, subclinical manifestations of eosinophil-moderated syndromes may be indexed by high-normal eosinophil counts, and hence information may be obtainable about the genetic basis of such syndromes by analysis of genetic determinants of eosinophil count within the normal range; that is, only using nominally healthy patients without overt eosinophilia. This has the advantage of a much larger potential study size. This suggested the leverage of an EGPA dataset with a large GWAS on EC in asymptomatic individuals.

Asthma and AAV

Asthma is an inflammatory disease of the bronchi, characterised by recurrent reversible obstruction of the airways. It is typically treated with long-term inhaled corticosteroids and symptomatically managed with beta-agonists. Asthma is frequently associated with eosinophilic airway infiltration, and is genetically associated with many immunomodulatory genes [Moffatt et al., 2010]. It is a common disease, with a prevalence of approximately 5% in the UK [Anderson et al., 2007] and substantial mortality. Given the asthmatic prodrome of EGPA, and the commonality of the phenotype, asthma was a clear option for leverage for genetic discovery.

Another candidate for leverage was non-EGPA AAV, given the partly vasculitic nature of EGPA. AAV refers to a class of rare vasculitides all characterised by inflammatory cell infiltrates and vascular necrosis, generally involving small and medium-sized arteries. They are usually treated with immunosuppressive therapy, often cyclophosphamide. A GWAS on non-EGPA AAV was conducted in parallel with the current study, providing another potentially useful dataset for leverage.

Genetics of EGPA

Prior to this study, data on genetic associations of EGPA was limited. Previous case-control studies [Vaglio et al., 2007, Wieczorek et al., 2008] found associations with HLA-DRB04 and DRB07, neither at genome-wide significance. This suggests a degree of heritability of EGPA, potentially with shared contributions from other autoimmune-associated loci.

The current study was designed as the first GWAS of EGPA, with the aim of developing understanding of the genetic basis. Because of the rarity of the phenotype, a relatively small cohort of 542 samples were able to be recruited.

3.3.2 Methods

Quality control

Quality control and the primary GWAS analysis were largely undertaken by my collaborators, although I helped in some areas. Notes on quality control are reproduced from the manuscript currently in preparation for this work.

599 individuals with a clinical diagnosis of EGPA were recruited from 17 centers in 9 European countries (UK, Ireland, Germany, Czech republic, Poland, France, Italy, Spain and Sweden). 9 individuals were excluded because of uncertain diagnosis. Genotype data for 6000 UK controls was obtained from the European Prospective Investigation of Cancer (EPIC) Consortium. Additional controls were also recruited and genotyped in parallel with the EGPA samples. 496 individuals with a history of asthma were excluded.

Samples with a sex mismatch, abnormal heterozygosity or proportion of missing genotypes >5%, or which were duplicated, were removed. SNPs were removed if they were monomorphic, had missing calls >2%, or deviated substantially from Hardy-Weinberg equilibrium (p-value $< 1 \times 10^{-6}$).

All quality control criteria for samples and SNPs were applied firstly to each batch of samples and secondly across the combined dataset. After combining samples, SNPs with differential missingness between cases and controls were removed (Fisher's exact test, FDR $< 5\%$ by Benjamini-Hochberg procedure). Principal components of the post-QC genotype calls combined with calls from 1000 Genome individuals (downloaded from http://www.1000genomes.org) were computed and, on the basis of the first three components, samples of non-European ancestry were excluded. After quality control, 542 EGPA cases and 6717 controls remained, with 543,639 genotyped SNPs.

Principal components were calculated for the genotype matrix including both cases and controls. Summary statistics were computed using the first 20 principal components as

covariates. The resultant summary statistics showed some residual inflation ($\lambda = 1.09$). In the primary GWAS analysis and the cFDR analysis, the unadjusted p-values were reported, along with indication of the SNPs which met an adjusted level of significance α equivalent to $p < 5 \times 10^{-8}$, using an analogous method to standard scaling of χ^2 scores [Devlin et al., 2001]:

$$\alpha = 2\Phi\left(-\sqrt{\lambda\Phi^{-1}\left(\frac{5\times10^{-8}}{2}\right)}\right) \approx 1.26\times10^{-8} \tag{3.4}$$

SNPs were included in each cFDR analysis if they met QC criteria for both the constituent studies. QC criteria for the EC and asthma GWASs can be found in the relevant papers ([Astle et al., 2016, Moffatt et al., 2010]). The GWAS on AAV included 609 cases, with the same control set and quality control procedures as used for the EGPA study. For the cFDR analysis, the MHC region was removed with wide boundaries (chr6:28.7-34.5 Mb, build hg19). Although imputation was performed for the GWAS analysis, only directly genotyped SNPs were used for cFDR. Sample sizes and available SNPs in the cFDR analyses are shown in table 3.7.

Table 3.7 Number of cases and controls for conditional phenotypes, shared controls and number of SNPs in cFDR analyses of EGPA

Analysis	Controls	Cases	Shared controls	N SNPs
EGPA|EC	173480		0	513,801
EGPA|Asthma	16110	10365	0	74776
EGPA|AAV	6717	609	6717	543,117

No adjustment for shared controls was necessary in the EGPA|Asthma and EGPA|EC analyses; the shared controls in the EGPA|AAV analysis led to a correlation of 0.08 between Z-scores for null SNPs. The parametrisation of the distribution of Z-scores for AAV found $(\pi_0, \sigma) = (0.71, 1.09)$.

Thresholds on cFDR were chosen to match FDR bounds with the GWAS analysis ($\alpha_2^{EGPA|\cdot}$; see equation 3.2.2) in the previous section, using the threshold adjusted for genomic inflation (equation 3.4) and noting that the GWAS had higher dimensionality than the cFDR analysis. Cutoffs are shown in table 3.8. Bounds are substantially higher than in table 3.2 due to a higher false-discovery rate bound in the EGPA GWAS than in the JIA GWAS.

3.3.3 Results

Q-Q plots showing observed and expected distributions of summary statistics for EGPA|EC and EGPA|asthma are shown in figure 3.2. I used adjusted p_{EGPA} values for the EGPA|AAV

Table 3.8 Thresholds on $\widehat{cFDR}$ for analysis of EGPA (FDR < 6.4×10^{-3} for all three analyses)

	$\alpha_2^{EGPA\|\cdot}$
EGPA|EC	2.8×10^{-3}
EGPA|Asthma	1.5×10^{-3}
EGPA|AAV	6.3×10^{-4}

plot. Inflation of summary statistics with conditioning is evident in the EGPA|Asthma and EGPA|EC analyses, but not in the EGPA|AAV analysis. Analysis of EGPA|Asthma yielded two SNPs reaching GWS by $\widehat{cFDR}$ not in regions reaching GWS for EGPA alone, shown in table 3.9. SNP rs11745587 is near the *IRF1/IL5* genes, in a region of chromosome 5 associated with CRO, UC, JIA, PSO and alopecia areata. IL5 (interleukin-5) is an important mediator of eosinophil activation [Sehmi et al., 1992], suggesting that variation in this region is a good candidate for mediating EGPA risk.

Table 3.9 Results for cFDR analysis of EGPA|Asthma

rsID	Chr.	Pos.	P(EGPA)	P(Asthma)	$\widehat{cFDR}$
rs11745587	5	131796922	1.10×10^{-06}	1.83×10^{-03}	1.05×10^{-04}
rs6454802	6	90814199	4.72×10^{-07}	2.16×10^{-03}	5.05×10^{-05}
rs1623646	10	9076230	1.49×10^{-05}	1.98×10^{-03}	6.05×10^{-04}

SNP rs6454802 is near *BACH2*, in a region of chromosome 6 with multiple autoimmune associations. SNP rs1623646 is in a relatively un-annotated region of chromosome 10 associated with RA [Okada et al., 2014]. The top EGPA SNP in the region, rs7898135 (not genotyped in the asthma dataset), had a p-value of 5.75×10^{-7}. In a chromatin conformation analysis, rs7898135 interacted with the *GATA3* gene in foetal thymus tissue and CD4+ cells in naive and non-activated states [Javierre et al., 2016, Schofield et al., 2016].

None of the SNPs found in the EGPA|Asthma analysis appeared to be strongly associated with asthma. However, other SNPs in the same regions were more strongly asthma-associated: rs1295686, near *IRF1/IL5*, had $p = 1.4 \times 10^{-7}$; rs4142967 near *BACH2*, had $p = 1.57 \times 10^{-5}$, and rs1242987, near rs1623646, had $p = 1.03 \times 10^{-4}$. This suggests that if asthma and EGPA are both associated at these regions, they are likely to have different causal architecture.

The EGPA|EC analysis found association with eight regions not reaching GWS by p-value for EGPA. These are shown in table 3.10. These findings replicated the associations found in the EGPA|Asthma analysis on chromosomes 5,6, and 10. SNP rs9290877 is in a region on chromosome 3 near *LPP*. SNP rs42041 is in a region on chromosome 7 associated with RA, near *CDK6*. SNP rs187564398 on chromosome 12 in a relative gene desert. SNP

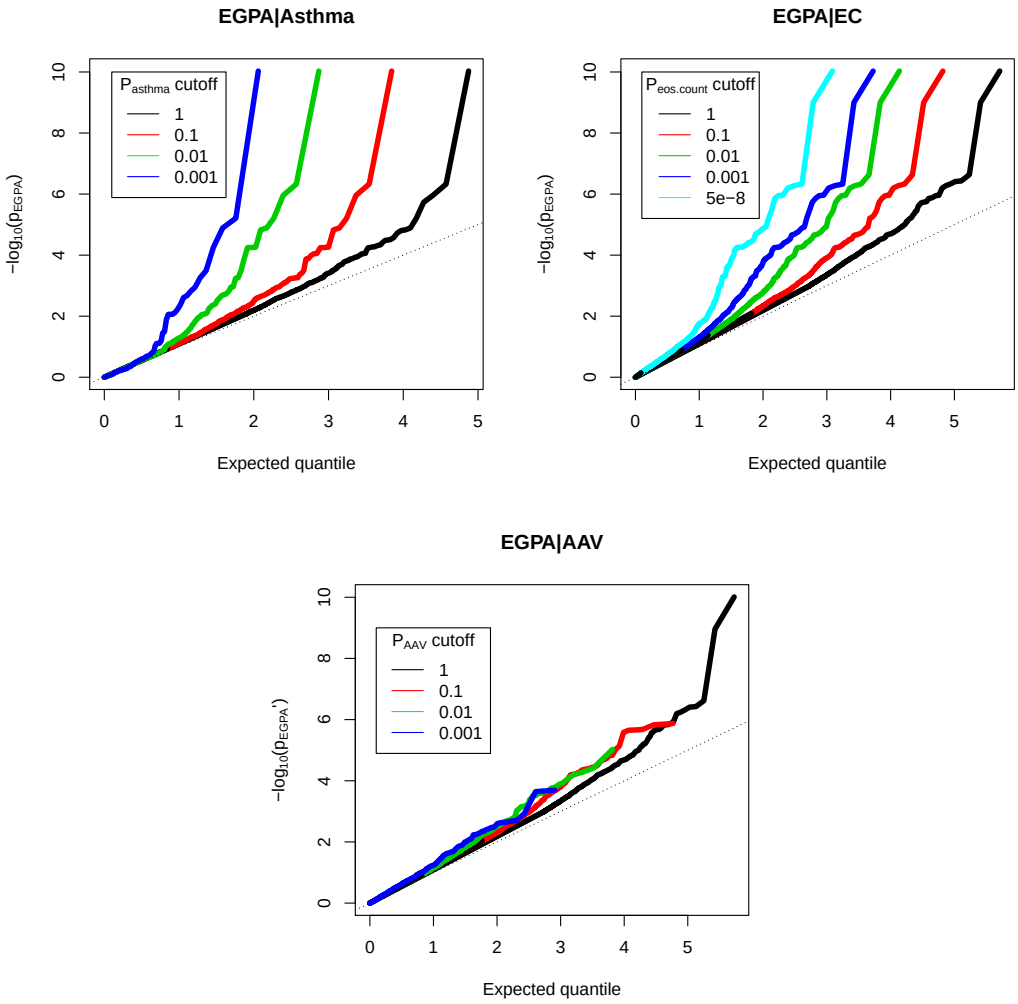


Fig. 3.2 Conditional Q-Q plots showing inflation of summary statistics for one phenotype conditioning on another (EGPA|Asthma on left, EGPA|EC on right, EGPA|AAV below). Note different cutoffs for p-value for the conditional phenotype.

rs2033784 on chromosome 15 is in a CRO and UC associated region near *SMAD3*. SNP rs4781047 on chromosome 16 is in a region associated with multiple autoimmune conditions near *CLEC16A* and *DEXI*.

Table 3.10 Results for cFDR analysis of EGPA|Asthma

rsID	Chr.	Pos.	P(EGPA)	P(EC)	$\widehat{cFDR}$
rs9290877	3	188442480	3×10^{-05}	9.06×10^{-14}	1.25×10^{-03}
rs11745587	5	131796922	1.1×10^{-06}	1.65×10^{-29}	1.62×10^{-04}
rs6454802	6	90814199	4.72×10^{-07}	1×10^{-18}	4.76×10^{-05}
rs42041	7	92246744	6.36×10^{-07}	4.56×10^{-08}	1.55×10^{-04}
rs1444782	10	9058671	3.53×10^{-06}	1.01×10^{-27}	2.95×10^{-04}
rs187564398	12	115934855	2.32×10^{-07}	6.94×10^{-03}	8.78×10^{-04}
rs2033784	15	67449660	4.35×10^{-05}	5.49×10^{-11}	1.92×10^{-03}
rs4781047	16	11318537	2.04×10^{-05}	7.78×10^{-14}	9.21×10^{-04}

No SNPs reached genome-wide significance by cFDR in the EGPA|AAV analysis, likely due to the relatively small size of the study for the conditional phenotype.

3.4 Discussion

This chapter extends chapter 2 by applying the cFDR method in new contexts and by using it in a new study to directly improve discovery of disease associations. These applications demonstrate the scope of the method, and explore the behaviour of the test statistic and associated false discovery rate in a wider variety of circumstances.

An important question which was pertinent in this chapter was why a complex procedure such as cFDR should be performed for leverage, as opposed to simply considering SNPs reaching a pre-set threshold for association for the conditional phenotype and determining association with the principal phenotype using a Bonferroni correction or similar (such as the analysis in [Plagnol et al., 2011]). Indeed, conditioning on genome-wide significance for eosinophil count in the EGPA|EC analysis may lead to similar results, with an easier interpretation. In general, however, the cFDR should in general be a more powerful tool for SNP discovery, despite being more complicated. The main reason for this is that an analysis conditioning on a single threshold disregards most of the information the conditional GWAS may contribute. This is discussed further in chapter 7, section 7.1.4

A second reason is that the choice of single threshold to condition on is difficult. If it is chosen in a truly *a priori* way, before any data is observed, then the choice is somewhat arbitrary; the genome-wide significance threshold of $p < 5 \times 10^{-8}$ is relatively conservative,

especially if the conditional GWAS is small. If the single threshold is chosen after observing the data, as probably happens often in practice, this introduces bias to the analysis. By not requiring any predefined threshold on association in the conditional phenotype, the cFDR circumvents these problems.

Conditioning on a single p-value threshold for the conditional phenotype may be appealing in its simplicity. It is still, however, susceptible to an inflated false-positive rate from that suggested by the Benjamini-Hochberg procedure if controls are shared, and appropriate adjustments should be made in this case. A recent example of such an adjustment was made in a GWAS for T1D [Onengut-Gumuscu et al., 2014] levering on celiac disease; while previous analyses had declared SNPs associated with T1D with $p_{T1D} < 1 \times 10^{-4}$ if $p_{celiac} < 5 \times 10^{8}$, the presence of shared controls required tightening the threshold on p_{T1D} to $p_{T1D} < 1 \times 10^{-5}$ to maintain a Bayesian posterior probability of T1D association greater than 0.9. Adaptations of the methodology in chapter 2, sections 2.2.3, 2.4.3 could be used in a similar way.

Another subtle area in which shared controls may improperly suggest inflation is in conditional Q-Q plots; it is possible to use either raw p-values or adjusted p-values p'. In general, I consider it appropriate to replace the p-values for the principal phenotype with the expected quantile (see chapter 2, sections 2.2.3, 2.4.3 for details) when controls were shared.

A major consideration in the analysis of EGPA is the choice of which phenotypes (and more generally, how many phenotypes) to use for leverage by cFDR. Different diseases may share different causative processes with the disease of interest, suggesting that leverage on a greater number of phenotypes could aid in the discovery of more associations than levering on a small number of phenotypes. However, leverage on a large number of phenotypes necessitates a complex multiple-testing correction; one example of this is discussed in section 2.4.8, although I have not tested this procedure on a larger scale ($\gg 10$ conditional phenotypes). If phenotypes used for leverage only minimally improve the power of the analysis, then the need to control for multiple testing may outweigh the advantage gained from including such phenotypes. This could be resolved with the use of a prior on the expected contribution of each conditional phenotype on the analysis (derived from phenotype ontologies [Robinson et al., 2008] or similar) but I did not explore this approach in this context.

If there is no pleiotropy between phenotypes, then (considering observed p-value pairs (p_i, p_j) as iid observations of a bivariate random variable (P_i, P_j) and using the notation in chapter 2, section 2.2.2) we have

$$E(\widehat{cFDR}(p_i|p_j)) = E(\widehat{uFDR}(p_i) = \frac{p_i}{Pr(P_i \leq p_i)} \tag{3.5}$$

since $Pr(P_i \leq p_i)$ is independent of p_j. If there is pleiotropy, then in general

$$E(\widehat{cFDR}(p_i|p_j)) \neq E(\widehat{uFDR}(p_i)$$

and the expected ranking of SNPs by p-value (or $\widehat{uFDR}$) will differ from the expected ranking of SNPs by $\widehat{cFDR}$.

This leads to a second important point arising from this chapter. A practical consideration in the analysis of cFDR is the question of which SNP to report as the 'top SNP' in a region declared associated by cFDR - the SNP with the best $\widehat{cFDR}$ value, or the SNP with the best p-value for the principal phenotype. In this chapter and chapter 2, I generally took the former option, following the first paper in the field [Andreassen et al., 2013]. However, in consideration of the results from the EGPA|Asthma analysis, in which the SNP rs7898135 had a markedly lower p-value for EGPA association than did the SNP rs1623646 (which had the lowest $\widehat{cFDR}$ score in the region), I would recommend for future analysis that both SNPs be reported, as well as the top SNP in the region for the conditional phenotype. The cFDR method is only able to find associations on the scale of regions (LD-blocks) of the genome, and different diseases may have different patterns of causal variants in the same region. Indicating only the variant with lowest $\widehat{cFDR}$ may understate the association of the principal and conditional phenotype with the region. Furthermore, the expected ranking of SNPs by observed p-value may differ from the ranking by true effect size for technical reasons (for instance, poor imputation quality) which may also lead to errors if the 'top' SNP in a region is chosen on the basis of minimum cFDR score.

A GWAS and follow-up analyses can be considered to be attempting to find two things: firstly, the subset of variants in the genome associated with the trait, and secondly, the genetic architecture' of the trait; specifically the relative effect sizes of all causal variants. Because of the changed expected ordering implied by equation 3.6, the cFDR procedure disturbs the observed genetic architecture of the phenotype on both single-region and genome-wide scales. In a standard GWAS, effect sizes can be recovered from p-values and observed MAF, and more specifically, for variants of a given AF, the effect size is monotonic to the p-value. Moreover, if a single variant in an LD block is causal, the probability that the causal variant has the lowest p-value tends to 1 as sample sizes tend to infinity.

In cFDR analyses with pleiotropy between phenotypes, this is not the case: $E(\widehat{cFDR})$ is responsive to the p-value for both the principal and the conditional phenotype. If the set of causal variants in a region differ between the principal and conditional phenotypes, then the minimum cFDR value will not be monotonic to the p-value, and even with a single causal

variant for each phenotype will not have minimal expected value at the causal variant for the principal phenotype. The same consideration holds across the whole genome; unless the genetic architectures of the conditional and principal phenotype are identical, the cFDR will not generally order variants of equal MAF according to their relative effect sizes in the way that a standard GWAS analysis would. This point is discussed further in chapter 7, section 7.1.1.

In several of the cFDR analyses presented, some SNPs reaching genome-wide significance by $\widehat{cFDR}$ when they did not by p-value did not appear to have a low enough p-value for the conditional phenotype to justify the apparent improvement in evidence for association (for example, SNPs rs7808122 and rs34132030 in the JIA|T1D analysis (table 3.3) and SNPs rs2364480, rs79893749 and rs4755450 in the JIA|RA analysis (table 3.4)). This is an effect which can occur even when no pleiotropy is present. Although the expected values of $\widehat{cFDR}$ and $\widehat{uFDR}$ are the same (equation 3.5), and hence their expected order is the same (and the same as the order of expected p-values), the deviation from this ordering differs between $\widehat{cFDR}$ and $\widehat{uFDR}$. If thresholds γ_1, γ_2 are chosen for $\widehat{cFDR}$ and $\widehat{uFDR}$ in order to control the false-discovery rate at the same level between procedures, some variants with $\widehat{uFDR} < \gamma_1$ may 'move' such that $\widehat{cFDR} > \gamma_2$ and vice versa, despite the overall FDR bound being maintained. The chances of a SNP moving 'up' or 'down' this way are approximately equal. However, if $\alpha_1^{(\cdot|\cdot)}$ is used as a $\widehat{cFDR}$ threshold, no SNPs can move 'out' of genome-wide significance by $\widehat{uFDR} < \gamma/p_i < 5 \times 10^{-8}$, although SNPs can move in. For this reason, I consider that the cutoff $\alpha_2^{(\cdot|\cdot)}$ is a more appropriate approximation of GWS by p-value.

The EGPA result suggests that EGPA could be primarily driven by high eosinophil count, or caused by some external process which also causes high-normal count in healthy individuals. An interesting subsequent analysis if possible would be to consider whether the disease has any evidence of causative variants not associated with physiological eosinophil count.

The idea of levering an association analysis for a rare disease using a much larger dataset for an associated trait is an interesting avenue for future research. One such avenue may be to lever on a meta-analysis of multiple diseases, making use of shared associations. This could lead to difficulty discovering variants with both deleterious and protective effects across the disease spectrum (which are common; see [Cotsapas and Hafler, 2013]) due to 'averaging' across diseases, so meta-analysis by p-value may be necessary. This is discussed further in chapter 7, section 7.2.1.

The results presented here give information on the set of genetic associations of the diseases in question additional to that attainable by p-value analysis alone, although as

expected, the ability to discover new associations is limited if the dataset used has already been superseded in size or power by subsequent studies. In particular, these results highlight the pervasive sharing of genetic associations between autoimmune phenotypes. One striking point is the effectiveness of conditioning on a non-disease trait, namely eosinophil count in healthy individuals, in understanding disease aetiology. The dataset of blood cell traits [Astle et al., 2016] is a particularly exciting one in this respect. Other BioBank data sets (eg [Canela-Xandri et al., 2017]) will also be interesting for leverage purposes.

Chapter 4

Two-stage testing with shared controls

4.1 Introduction

High-dimensional case-control studies have become a mainstay of investigation of pathophysiology in complex diseases and traits. An important part of their analysis is the process of replication [Wason and Dudbridge, 2012], in which the results of a high-dimensional study are used to inform the design of a second study at a subset of the original variables, with a joint analysis used to determine overall association.

Replicating studies in this way has the advantage of increasing the effective study sample sizes without requiring measurement of all variables in all samples. It also serves to protect against false-positives due to systematic errors in the original datasets, by re-testing association in a second nominally independent dataset.

Replication has a significant cost, and can require large numbers of samples, especially when associated variables have small effects (ie [Fuchsberger et al., 2016]). There is therefore a need to minimise the number of additional samples which need to be analysed. This paper presents a method to perform replication by combining controls in both the original 'discovery' and second 'replication' datasets, potentially reducing the number of new samples required. Shared-control approaches can improve study efficiency in many related applications in which studies are compared [Lin and Sullivan, 2009, Han et al., 2016a, Bhattacharjee et al., 2012, Zaykin and Kozbur, 2010, Liley and Wallace, 2015, Fortune et al., 2015].

Results from original and replication datasets for which some or all controls are shared cannot be directly compared due to the correlation between test statistics directly resulting from shared controls even under the null hypothesis [Bhattacharjee et al., 2012]; use of the same thresholds in a shared-control design as used in an independent-controls design

will lead to higher type-1 error rates. In this chapter I demonstrate a simple adaptation to a standard design to account for the changed correlation structure and retain control of type-1 error rate, only requiring a change to one p-value threshold.

The action of sharing control samples results in a different spectrum of sensitivity to confounding in study groups. It necessitates a sacrifice of type-1 error rate control in variables affected by confounding in the discovery-phase control cohort, but improves type-1 error rate control in variables affected by confounding in the replication-phase control cohort. Performance is largely equivalent to an independent-controls design for variables affected by confounding in either case cohort.

The new spectrum of false positive rates can be advantageous in circumstances where control samples in the replication cohort are less well-ascertained than those in the discovery cohort. This may be the case in studies on degenerative disease, where control ascertainment is generally uncertain, and population-sourced controls may be used for replication. The shared-control design can reduce power losses from mis-specified controls in the replication cohort, as well as reducing false-positive rates caused by confounding in the cohort.

When used with shared cases instead of controls, this method can be adapted to a 'partial replication' procedure where only a new control set is used. Although not equivalent to a full replication in an independent dataset, the procedure enables improvement in type-1 error rates and control over confounding. This is applicable in studies on rare traits, where all available samples need to be included in the discovery analysis for adequate power.

Throughout this chapter I use GWAS terminology (SNPs, allele frequency, variants etc) although the method is applicable to any high-dimensional case control study. 'Controls' will be considered to generally be samples unaffected by a disease or trait of interest, although the method can be applied with case/control labels swapped, or applied to comparisons between subgroups of a case group. Asymptotic analytical results are established where possible, but all type 1/type 2 error rates are readily tractable empirically to good accuracy given study sizes and proposed p-value thresholds, and a tool is provided to do this at https://wallacegroup-liley.shinyapps.io/replication_shared/.

4.2 Results

4.2.1 Overview of method

Assume a GWAS dataset of a set of cases C_1 and controls C_0 used in a 'discovery' phase of a GWAS or similar study, and corresponding sets of cases and controls C_1', C_0' in the replication

phase. Further assume that C_0 and C_1 are genotyped at a set of SNPs S and C_0', C_1' at a set $S' \subseteq S$.

For each SNP designate μ_1, μ_0, μ_1', μ_0' as the population minor allele frequency in the corresponding group, and m_1, m_0, m_1', m_0' as the observed allele frequency (so $E(m_i) = \mu_i$). Designate two null hypotheses; $H_0^{\cup} : (\mu_1 = \mu_0) \cup (\mu_1' = \mu_0')$ and $H_0^{\cap} : (\mu_1 = \mu_0 = \mu_1' = \mu_0')$, noting that $H_0^{\cap} \Rightarrow H_0^{\cup}$. In a typical conservative GWAS approach, we seek to test against $H_0^{\cup}$, since $\mu_1 \neq \mu_0$ or $\mu_1' \neq \mu_0'$ may hold at non-disease associated SNPs due to confounding in the original or replication studies respectively.

A typical two-stage genetic testing procedure [Wason and Dudbridge, 2012], which I will refer to as method A, begins by comparing genotypes of C_1 and C_0 at SNPs S generating p-values p_d (discovery). A subset S' of SNPs reaching putative significance level $p_d < \alpha$ are genotyped in C_0' and C_1', with genotypes compared to generate p-values p_r (replication stage). Finally, genotypes are compared between $C_0 \cup C_0'$ and $C_1 \cup C_1'$ at SNPs S' to generate p-values p_m (meta-analytic stage). SNPs are designated as 'hits' if $p_d < \alpha, p_r < \beta, p_m < \gamma$ for some β, γ, and all effects have the same direction.

The main modification proposed in this chapter, denoted as method B, differs at the replication stage in that C_1' is compared with $C_0 \cup C_0'$ at S' instead of just C_0' (figure 4.1). The p-values resulting from the modified replication stage are termed p_s, and the criterion to designate a hit changed to $p_d < \alpha, p_s < \beta^*, p_m < \gamma$, with all effects in the same direction. The threshold β^* is chosen to conserve type-1 error rate between methods (see section 4.4 and appendix B.1.1).

A second modification, denoted method C, combines C_0 and C_0' at both the discovery and replication phase (figure 4.1). This is analogous to a situation in which only a single control cohort is available, and a choice must be made to split it between discovery and replication procedures or to use it for both. In this case, $C_0 \cup C_0'$ is compared with C_1 at SNPs S in the discovery phase to produce p-values p_c, then $C_0 \cup C_0'$ is compared with C_1' at SNPs S' at the replication phase and compared with $C_1 \cup C_1'$ at the meta-analytic stage to produce p-values p_s and p_m as before. A hit is determined by $p_d < \alpha, p_s < \beta^{\perp}, p_m < \gamma$, with all effects in the same direction. Again, $\beta^{\perp}$ is chosen to maintain the type-1 error rate between methods.

4.2.2 General properties

For SNPs in $H_0^{\cap}$, the overall type-1 error rate is conserved between methods by the definition of β^*, $\beta^{\perp}$ (equation 4.2) at a level P_0. I show in appendix B, section B.1.2 that $\beta > \beta^* > \beta^{\perp}$. For SNPs in $H_0^{\cup} \setminus H_0^{\cap}$ the type-1 error rates differ between methods. Such SNPs may be

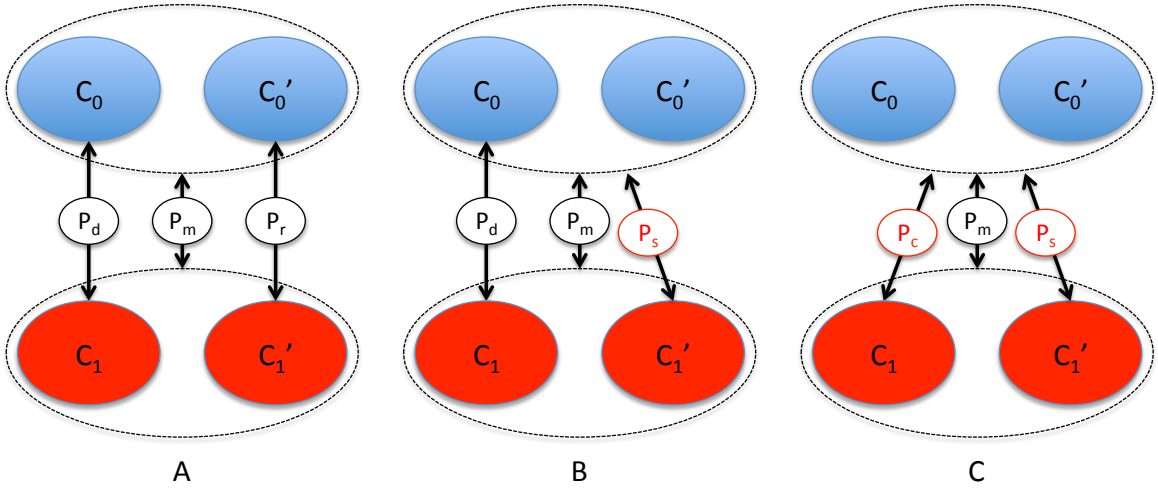


Fig. 4.1 Diagram of methods A, B, and C. Method B differs by comparing C_1' to pooled C_0 and C_0' at the replication stage to generate p-value P_s instead of P_r. Method C also pools controls at the discovery phase, comparing C_1 to pooled C_0 and C_0' to generate p-values P_c instead of P_d. A 'hit' is declared in method A if $P_d < \alpha$, $P_r < \beta$, $P_m < \gamma$, in method B if $P_d < \alpha$, $P_s < \beta^*$, $P_m < \gamma$ and in method C if $P_c < \alpha$, $P_s < \beta^\perp$, $P_m < \gamma$.

characterised by the group(s) amongst C_0, C_1, C_0', C_1' in which their expected MAF is aberrant from the expected MAF in the population which the group ostensibly represents. 'Aberrance' is taken to mean an incorrect expected value from systematic measurement error or uncorrected confounding, rather than random deviance around a correct expected value.

Bounds on type-1 error rates with aberrance in each group are shown in table 4.1. Methods B and C necessitate sacrificing bounds on error rates with aberrance in C_0 and C_0,C_0' respectively. The bound on error with aberrance in C_1' improves through methods A-C. I show in section 4.4.5 that the type-1 error with aberrance in C_0' decreases from methods A to B, and the error with aberrance in C_1' increases from A through C, although the upper bound is the same for both.

Table 4.1 Upper bounds on type 1 error rates with aberrance in cohorts, noting that $\beta > \beta^* > \beta^\perp$

	Aberrant cohort				
	None	C_0	C_0'	C_1	C_1'
M. A	P_0	β	α	β	α
M. B	P_0	1	α	β^*	α
M. C	P_0	1	1	$\beta^\perp$	α

4.2.3 Simulation

I analysed the power difference between methods B and A systematically across a range of values of (n_0, n_1, n_0', n_1'). I compared both average power difference and maximum power difference (see section 4.4.4). Figure 4.2 shows power difference at various study sizes for typical α, β, γ values ($\alpha = 5 \times 10^{-6}$, $\beta = 5 \times 10^{-4}$, $\gamma = 5 \times 10^{-8}$) and minor allele frequency 0.1. The difference is typically highest when the ratio of controls to cases is high in the discovery cohort and low or equal in the replication cohort, and the number of cases in the discovery cohort is larger than the number in the replication cohort. Power to detect SNPs in H_1 is typically highest in method C, second-highest in method B, and lowest in method A.

4.2.4 Recommended applications

To demonstrate areas where this approach is applicable, several examples are constructed or sourced from the GWAS field in which the procedure of sharing controls or cases will improve power or type-1 error profile of the two-stage testing procedure or enable some form of orthogonal replication to be performed.

Assumptions

In order to use method B or C, it must be assumed that cohort C_0 and C_0' are sampled from similar enough populations to be comparable to C_1 and C_1' (possibly with the inclusion of strata or covariates in the genetic risk model). An important caveat of methods B and C is sacrifice of control over errors arising from aberrance in C_0 (method B) or $C_0 \cup C_0'$ (method C), so an assumption must be made that variables affected by confounding or measurement error in these cohorts are understood to be distinguishable from true associations by quality-control measures only.

Post-hoc assessment of all putative hits should be performed to check for genotyping errors [Anderson et al., 2010] and assess whether the hit could have arisen from aberrance in C_0.

Conventional GWAS

Method B is applicable in several cases in large conventional GWAS, particularly when then ratio of controls to cases in the discovery cohort is larger than that in the replication cohort. In a relatively recent GWAS on rheumatoid arthritis [Stahl et al., 2010] with comparable sample populations for discovery and replication cohorts, method B could be used to attain

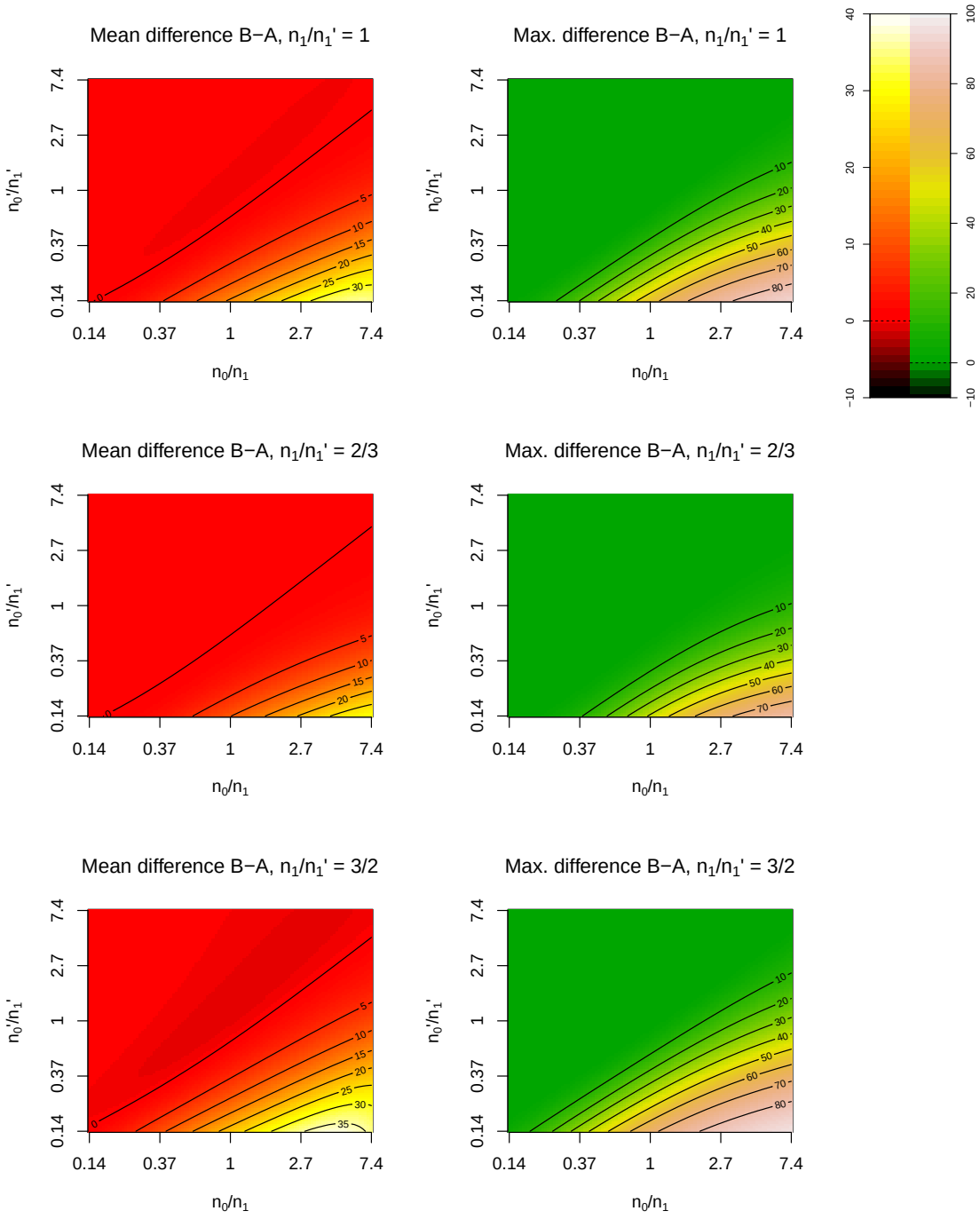


Fig. 4.2 General power differences (%) between methods A and B. Mean power difference is taken as the integral of power difference between methods B and A (see methods section) over $\mathbb{R}$ with respect to log-odds ratio. In all cases, 20 000 samples are used overall for a SNP with MAF 0.1, with cutoffs $\alpha = 5 \times 10^{-6}$, $\beta = 5 \times 10^{-4}$, $\gamma = 5 \times 10^{-8}$. Mean power difference is determined as the integral of the power difference with respect to the log-odds ratio over the real line.

greater power than method A for a fixed type-1 error rate. Assuming that summary statistics are well-approximated by binomial tests of allelic differences (so covariates and strata used in computation of summary statistics have only small effects), the improvement in power is around 4% for SNPs with an odds-ratio of 1.3, MAF 0.1, and is positive across all odds ratios. More than 2000 additional controls in C_0' would be needed to increase power by this amount (figure 4.3a).

Small power advantages such as this may make minimal difference in a single study, although since they require no extra cost, are worth attaining if possible. The power of method B is generally considerably higher than method A when $n_0 > n_1$ and $n_0' \approx n_1'$, which corresponds to a scenario in which cases are comparatively harder to recruit, but control recruitment and genotyping also carries cost (in that minimising the number of controls needed for replication is of interest). Power advantages may be more substantial in some cases; for example, a study with $(n_0, n_1, n_0', n_1') = (15000, 5000, 5000, 5000)$, method B enables a power increase of up to 8% (Figure 4.3b). To achieve comparable performance with method A, around 2000 additional controls would be necessary in the replication cohort. Method B with $(n_0, n_0') = (15000, 5000)$ is also more powerful than method A would be if controls were divided equally between C_0 and C_0' (see Figure 4.3b).

Difficult control ascertainment

An important application of the method presented in this paper is in studies for which 'control' samples are expensive or difficult to ascertain. This is often the case in comparative studies between disease subtypes. In such studies, sharing controls can improve power substantially, especially if a proportion of samples in the replication cohort are falsely assigned to the control cohort (see methods section).

An international GWAS on fronto-temporal dementia in 2014 [Ferrari et al., 2014] is an example in which sharing controls may be beneficial. The study had sample sizes $(n_0, n_1, n_0', n_1') = (4308, 2154, 5094, 1372)$. Control samples in the discovery phase were assessed for current neurological disease, and were used in previous studies on Parkinson's disease, indicating a high degree of reliability. Control samples in the replication phase were collected from the same geographic distribution as cases, but were not explicitly used in previous neurological studies, suggesting better control ascertainment amongst the discovery cohort.

In this study, sharing controls could allow for a more strongly-ascertained control cohort, and reduce the effects of confounders affecting C_1'. At typical values $\alpha = 1 \times 10^{-4}$, $\beta = 1 \times 10^{-3}$, $\gamma = 5 \times 10^{-8}$, power is nearly equivalent between the two methods assuming all

controls are genuine. However, with 10% mis-ascertainment in C_1', the power advantage of method B is up to 5%. Given the near-identical distribution of cases in the discovery and validation cohort, cases could alternatively be shared, leading to a power increase of up to 6%.

4.2.5 Prospective study design

Studies may be planned and powered with the assumption that samples may be shared. For certain restrictions on sample numbers, this can provide the potential for greater power than would be attainable by restricting to an independent-controls design. For instance, if we seek to validate hits on a GWAS with 10000 controls and 5000 cases, and can afford to genotype a further 10000 samples, power is higher after recruiting 4000 additional controls and 6000 additional cases and sharing controls than can be achieved from any independent-control study design (Figure 4.3d).

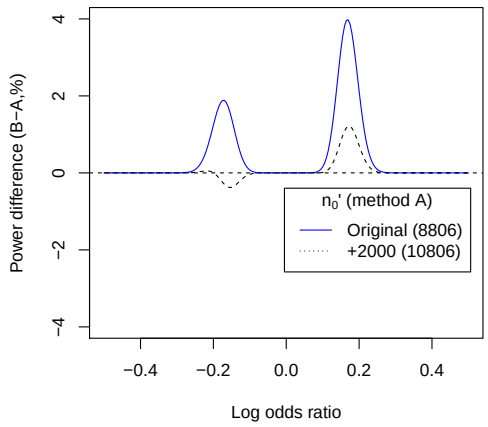
This may be a common scenario if controls are sourced from a known database rather than specifically recruited for the study.

Partial replication

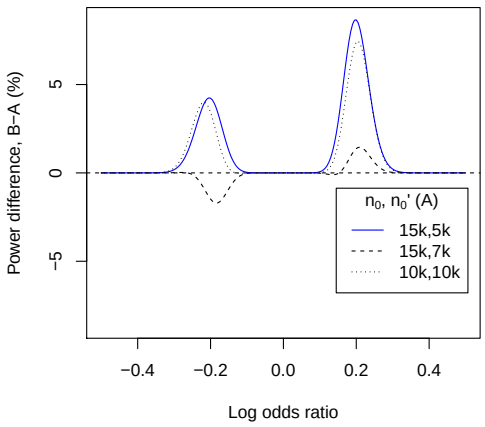
In circumstances where case recruitment is difficult, as in studies of rare diseases, an assessment of repeatability may be made by re-testing results from a discovery phase with a new control set only. This can enable the use of control cohorts which only partially match the case cohort.

In a GWAS on pemphigus vulgaris [Sarig et al., 2012], a rare disease primarily affecting individuals of Ashkenazi Jewish ethnicity, the discovery cohorts were sampled from Jewish populations, with age- and population- matched controls. Control cohorts were small ($(n_0, n_1, n_0', n_1') = (100, 400, 59, 285)$), potentially due to difficulty recruiting both ethnically- and geographically-matched controls.

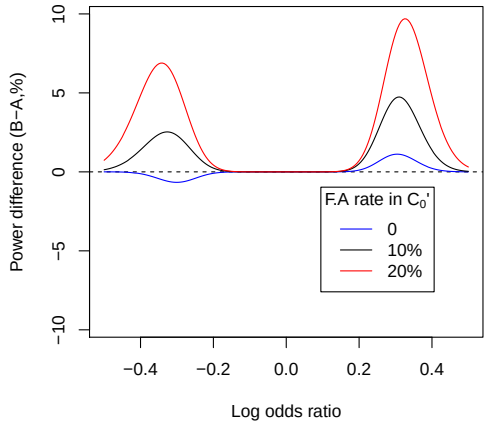
Method C could be used in this instance to enable a larger control set and greater power. If a control cohort of Ashkenazi individuals could be assembled without requiring geographic matching with the case set, it would be inappropriate to use as a sole control cohort against the existing case cohort, due to the potential for geographic confounding. However, such a cohort could be used as either C_0 or C_0' in method C, with the existing ethnically- and geographically-matched controls serving as the other cohort. In this way, the power advantage of the larger cohort could be used while maintaining control over potential aberrance in the larger control group.



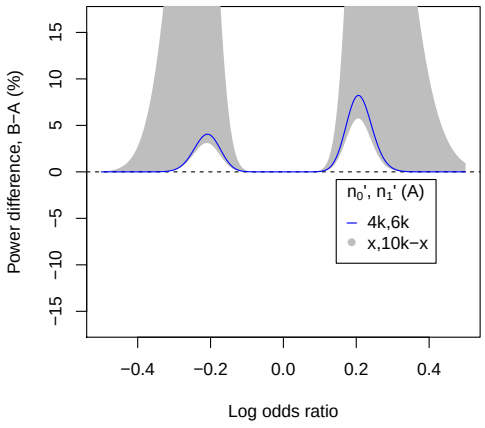
(a) Typical GWAS



(b) Example with $n_0 \gg n_0'$



(c) Example with incorrect ascertainment



(d) Prospective design

Fig. 4.3 Examples of comparison of power of methods A and B. In all panels, a positive odds ratio corresponds to a deleterious mutation and average MAF is 10%. The top two panels show comparisons of method B with n_0' fixed against method A with varying n_0'. Panel 4.3a has $(n_0, n_1, n_0', n_1') = (20169, 5539, 8806, 6768)$ (values from a GWAS on RA [Stahl et al., 2010]), and panel 4.3b $(n_0, n_1, n_0', n_1') = (15000, 5000, 5000, 5000)$. Both panels use $(\alpha, \beta, \gamma) = (5 \times 10^{-6}, 5 \times 10^{-4}, 5 \times 10^{-8})$. Panel 4.3c demonstrates the effect of false-ascertainment (F.A) in C_0'; when cases are mis-ascertained as controls. In this case, $(\alpha, \beta, \gamma) = (1 \times 10^{-4}, 1 \times 10^{-3}, 5 \times 10^{-8})$, reflecting values used in the paper [Ferrari et al., 2014]. Panel 4.3d demonstrates a prospective scenario with 10000 samples for replication. Method B with (n_0, n_1) as above, $(n_0', n_1') = (4000, 6000)$ is more powerful than any design using method A (grey region; $n_0' \in (1000, 9000)$; $n_1' = 10000 - n_0'$).

Method C enables computation of power and type-1 error rates, and comparison to alternative designs with cases split into smaller independent discovery and validation cohorts (method A). Testing a case cohort against two separate control cohorts is almost always more powerful for a fixed type-1 error rate than splitting the case cohort in two and performing method A (see supplementary figures B.1,B.2).

4.3 Discussion

This paper proposes a method to improve efficiency of data use in a replication procedure, adding to the body of methods for comparison of high-dimensional case-control studies. For many common study sizes, the method can reduce the cost of replication, or increase power of discovery. The adapted method is simple to apply, only requiring modification of a single association threshold. A standard replication procedure (or more general comparison of case-control studies) with independent control datasets does not make use of the information that expected values of variables in control datasets are, in principle, the same. In this way, the same dataset can in theory yield more information when controls are shared.

The most important caveat of these methods is the loss of systematic type-1 error rate control for null SNPs which are aberrant in C_0. Control of such errors must not be sacrificed entirely, but in some circumstances it may be satisfactory to assess such errors on a SNP-by-SNP basis. Such assessment is important and standard for all proposed GWAS hits under any method [The Wellcome Trust Case Control Consortium, 2007] in the interests of quality control. In method C, control over aberrance in C_0' is additionally lost; however, since this method is largely applicable when $C_0 \cup C_0'$ is a single homogeneous control (or case) cohort, there is no way that aberrance in the cohort can be systematically identified by comparison with other cohorts.

Somewhat better control of the type-1 error rate can often be achieved for SNPs with aberrance in C_1 or C_0'. This may incentivise the use of this method when confidence in the representativeness of these cohorts is low compared to that of C_0. The type 1 error rate is somewhat increased for SNPs with aberrance in C_1', although as it remains bounded by α, this increase is not a major problem.

The two-stage validation procedure is similar to a meta-analysis of the discovery and validation experiments, for which several adaptations to shared-control designs have been proposed [Lin and Sullivan, 2009, Han et al., 2016a]. However, there are several important distinctions which necessitate an alternative approach in this case. Firstly, not all variables are measured in the second (replication) study; we are restricted to analysis of variables

reaching a given observed effect size. Secondly, the studies to be 'meta-analysed' are not complete, in the sense that there may be residual confounding; a strong effect size in the meta-analysis alone is not adequate evidence for association and some level of association (with consistent direction) is additionally required in both constituent studies.

The method is inapplicable when replication is performed on cohorts from completely distinct geographic groups, although there can be some difference in geographic distribution between control sets if this is controlled for in computing summary statistics. The method is most applicable when control groups are sampled from similar populations and genotyped on similar platforms.

The widespread discoveries of the GWAS field have led to corresponding increases in complexity of phenotypic definitions, with ever-finer delineations of disease types of ever-rarer prevalence. The genetic analysis of such complex phenotypes is necessarily comparative; there is little use understanding the genetics of a rare disease subtype except in the context of the genetics of the disease in general. Such analyses necessitate GWAS and other comparative studies between rare phenotypic types [Liley et al., 2016], with 'controls' meaning the better-characterised disease sub-phenotype in this sense, as well as between cases and controls. Rare disease subtypes are often afflicted with ascertainment difficulties, leading to varying degrees of expected aberrance in disease cohorts. Within this paradigm, the applicability of this method is likely to expand.

4.4 Methods

4.4.1 Definitions

Denote z_x for $x \in \{d,r,s,m,c\}$ as the signed z-score ($\pm\Phi^{-1}(p_x/2)$) corresponding to p_x, and z_x for $x \in \alpha, \beta, \beta^*, \gamma$ as the positive corresponding threshold $-\Phi^{-1}(x/2)$, where Φ, Φ^{-1} are the standard normal CDF and quantile functions. Other than (z_d, z_r), all pairs of z-scores are correlated under $H_0^\cap$, with correlation estimable from sample sizes or empirically if covariates are used (appendix B, section B.1.1). Denote ρ_{xy} as the correlation between z_x and z_y, $(x,y) \in \{d,r,s,m\}^2$, and set

$$\Sigma_A = \text{var}\left((z_d\, z_r\, z_m)^t\right) \qquad \Sigma_B = \text{var}\left((z_d\, z_s\, z_m)^t\right) \qquad \Sigma_C = \text{var}\left((z_c\, z_s\, z_m)^t\right) \tag{4.1}$$

For $i \in \{d,r,s,m,c\}$ define $\zeta_i = E(z_i)$, where the expectation is conditional on the SNP in question. For SNPs in $H_0^\cap$ we have $\zeta_i \equiv 0$, but this may not hold for SNPs in $H_0^\cup \setminus H_0^\cap$. Note that the values ζ_i are proportional to the corresponding log-odds ratios for SNPs with fixed MAF. Define R_A, R_B, R_C as the false-positive rates for a SNP of interest in methods A, B and C respectively.

4.4.2 General type 1 error rate

Define $N_\Sigma(\mathbf{z})$ as the PDF of the multivariate normal with mean 0 and variance Σ at $\mathbf{z}$. The values β^*, $\beta^\perp$ are chosen to satisfy

$$\begin{aligned} 2\int_{z_\alpha}^{\infty}\int_{z_{\beta^*}}^{\infty}\int_{z_\gamma}^{\infty} N_{\Sigma_B}\left((z_d\ z_r\ z_m)^t\right) dz_m dz_r dz_d &= 2\int_{z_\alpha}^{\infty}\int_{z_{\beta^\perp}}^{\infty}\int_{z_\gamma}^{\infty} N_{\Sigma_C}\left((z_d\ z_r\ z_m)^t\right) dz_m dz_r dz_d \\ &= 2\int_{z_\alpha}^{\infty}\int_{z_\beta}^{\infty}\int_{z_\gamma}^{\infty} N_{\Sigma_A}\left((z_d\ z_r\ z_m)^t\right) dz_m dz_r dz_d \\ &= Pr(p_d < \alpha, p_r < \beta, p_m < \gamma | H_0^\cap) \end{aligned} \tag{4.2}$$

thus conserving the type 1 error rate (denoted P_0) against $H_0^\cap$ between methods (Figure 4.4). If no threshold is used on p_m (ie, $\gamma = 1$), then β^*, $\beta^\perp$ satisfy

$$Pr(p_d < \alpha, p_s < \beta^* | H_0^\cap) = Pr(p_c < \alpha, p_s < \beta^\perp | H_0^\cap) = Pr(p_d < \alpha, p_r < \beta | H_0^\cap) = \alpha\beta \tag{4.3}$$

since $z_d \perp\!\!\!\perp z_r | H_0^{\cap}$. The definition of $\beta^*, \beta^\perp$ from equation 4.2 will be considered a generalisation of the definitions from 4.3, with results established first for β^*, $\beta^\perp$ defined as per definition 4.3, and extended where possible to definition 4.2.

I show in appendix B, section B.1.2 that for β^*, $\beta^\perp$ defined as per definition 4.3 we have

$$\lim_{z_\alpha \to \infty} \frac{z_{\beta^*}}{\sqrt{1-\rho_{ds}^2} z_\beta + \rho_{ds} z_\alpha} = 1 \tag{4.4}$$

$$\lim_{z_\alpha \to \infty} \frac{z_{\beta^\perp}}{\sqrt{1-\rho_{cs}^2} z_\beta + \rho_{cs} z_\alpha} = 1 \tag{4.5}$$

approaching from above, so

$$z_{\beta^*} > \max\left(z_\beta, \sqrt{1-\rho_{ds}^2} z_\beta + \rho_{ds} z_\alpha\right) \tag{4.6}$$

$$z_{\beta^\perp} > \max\left(z_\beta, \sqrt{1-\rho_{cs}^2} z_\beta + \rho_{cs} z_\alpha\right) \tag{4.7}$$

As defined by equation 4.3, z_{β^*}, $z_{\beta^\perp}$ are also asymptotically linear in z_α, z_γ, z_β as the former two tend to ∞, with some constraints (appendix B, section B.1.2), although the limit does not necessarily approach from above. For both definitions, $\beta^\perp < \beta^* < \beta$ (appendix B, section B.1.2)

4.4.3 Study sizes, odds ratios and allele frequencies

Consider a study with n_0 controls and n_1 cases, with underlying allele frequencies μ_0 and μ_1 and observed allele frequencies m_0, m_1 in controls and cases respectively. Let Z be a signed Z-score derived from a GWAS p-value against the null hypothesis $\mu_0 = \mu_1$. To first order,

$$E(Z) = \sqrt{\frac{2n_0 n_1}{n_0 + n_1}} \frac{\mu_1 - \mu_0}{\sqrt{\bar{\mu}(1-\bar{\mu})}} \tag{4.8}$$

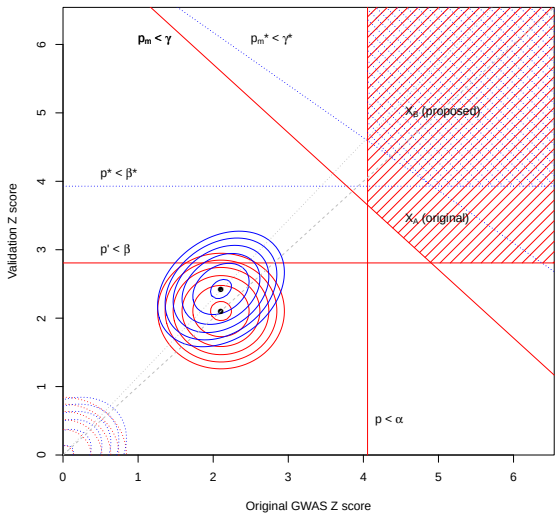


Fig. 4.4 Replication with shared controls. Red and blue shaded areas are regions where a pair of observed Z scores are deemed a 'hit' in the $(+,+)$ quadrant under method A/B respectively. The value z_m is almost linearly dependent on (z_d, z_r) and on (z_d, z_s) (appendix B, section B.1.1). Solid red/blue ellipses indicate contours of the distribution of observed Z scores for a typical disease-associated SNP under methods A and B, and dashed ellipses indicate contours for a SNP in $H_0^{\cap}$.

where $\bar{\mu} = \frac{n_0\mu_0+n_1\mu_1}{n_0+n_1}$. Hence (approximately)

$$\zeta_d = \sqrt{\frac{2n_0n_1}{n_0+n_1}}\frac{\mu_1-\mu_0}{\sqrt{\bar{\mu}(1-\bar{\mu})}} \qquad \zeta_r = \sqrt{\frac{2n'_0n'_1}{n'_0+n'_1}}\frac{\mu'_1-\mu'_0}{\sqrt{\bar{\mu}(1-\bar{\mu})}}$$
$$\zeta_s = \sqrt{\frac{2(n_0+n'_0)n'_1}{n_0+n'_0+n'_1}}\frac{\mu'_1-\frac{\mu_0n_0+\mu'_0n'_0}{n_0+n'_0}}{\sqrt{\bar{\mu}(1-\bar{\mu})}} \qquad \zeta_c = \sqrt{\frac{2(n_0+n'_0)n_1}{n_0+n'_0+n_1}}\frac{\mu_1-\frac{\mu_0n_0+\mu'_0n'_0}{n_0+n'_0}}{\sqrt{\bar{\mu}(1-\bar{\mu})}}$$
$$\zeta_m = \sqrt{\frac{2(n_0+n'_0)(n_1+n'_1)}{n_0+n'_0+n_1+n'_1}}\frac{\frac{\mu_1n_1+\mu'_1n'_1}{n_1+n'_1}-\frac{\mu_0n_0+\mu'_0n'_0}{n_0+n'_0}}{\sqrt{\bar{\mu}(1-\bar{\mu})}} \tag{4.9}$$

where $\bar{\mu}$ varies between definitions (though I generally will take it to be approximately constant). These formulae allow ζ_i to be estimated in empirical computations. The estimation of ζ_i is more complex if covariates or strata are used in the computation of z_i (appendix B, section B.1.1).

False ascertainment

In general, for a true association, $\mu_0 = \mu'_0$ and $\mu_1 = \mu'_1$. If some proportion κ of samples in C'_0 are incorrectly assigned and come from the case population, then $\mu'_0 = (1-\kappa)\mu_0 + \kappa\mu_1$. This lowers the absolute values of ζ_r, ζ_s and ζ_m, reducing the power to detect the SNP.

4.4.4 Empirical computations

The power to reject a SNP given ζ_d, ζ_r, ζ_s, ζ_c ζ_m in each method can be effectively estimated in terms of z_α, z_β, z_γ, z_{β^*}, $z_{\beta^\perp}$ by integrating over bivariate normals with covariance matrices Σ_A, Σ_B, Σ_C. Determination of these matrices is described in appendix B, section B.1.1. The probability of rejecting the null for a given SNP using method A is

$$\int_{z_\alpha-\zeta_d}^{\infty}\int_{z_\beta-\zeta_r}^{\infty}\int_{z_\gamma-\zeta_m}^{\infty} N_{\Sigma_A}\left((z_d\ z_r\ z_m)^t\right) dz_m dz_r dz_d$$
$$+\int_{z_\alpha+\zeta_d}^{\infty}\int_{z_\beta+\zeta_r}^{\infty}\int_{z_\gamma+\zeta_m}^{\infty} N_{\Sigma_A}\left((z_d\ z_r\ z_m)^t\right) dz_m dz_r dz_d \tag{4.10}$$

and using method B

$$\int_{z_\alpha-\zeta_d}^{\infty}\int_{z_\beta-\zeta_s}^{\infty}\int_{z_\gamma-\zeta_m}^{\infty} N_{\Sigma_B}\left((z_d\ z_s\ z_m)^t\right) dz_m dz_s dz_d$$
$$+\int_{z_\alpha+\zeta_d}^{\infty}\int_{z_\beta+\zeta_s}^{\infty}\int_{z_\gamma+\zeta_m}^{\infty} N_{\Sigma_B}\left((z_d\ z_s\ z_m)^t\right) dz_m dz_s dz_d \tag{4.11}$$

Matrix Σ_C is generally singular, and z_m can be written as

$$z_m = \frac{\rho_{cs}\rho_{sm} - \rho_{cm}}{\rho_{cs}^2 - 1} z_d + \frac{\rho_{cs}\rho_{cm} - \rho_{sm}}{\rho_{cs}^2 - 1} z_s = S(z_d, z_m) \tag{4.12}$$

allowing the probability of rejecting the null in method C to be written as a two-dimensional integral across the regions

$$X_1 = \{(z_d, z_c) : z_d > z_\alpha, z_c > S(z_d, z_m)\}$$
$$X_2 = \{(z_d, z_c) : z_d < -z_\alpha, z_c < -S(z_d, z_m)\}$$

using the covariance matrix $\Sigma'_C = \begin{pmatrix} 1 & \rho_{dc} \\ \rho_{dc} & 1 \end{pmatrix}$, as

$$\iint_{X_1 \cup X_2} N_{\Sigma'_C}\left((z_d\ z_c)^t\right) dz_d dz_c \tag{4.13}$$

If $\frac{n_0}{n_1} = \frac{n'_0}{n'_1}$, matrix Σ_A is singular (appendix B, section B.1.1), in which case $z_m = \rho_{dm} z_d + \rho_{vm} z_v$ and the expression above may be reduced to a two-dimensional integral over a more complex region in the same way (Figure 4.4). If Σ_B is nearly singular, the approximation

$$z_m \approx \frac{\rho_{ds}\rho_{sm} - \rho_{dm}}{\rho_{ds}^2 - 1} z_d + \frac{\rho_{ds}\rho_{dm} - \rho_{sm}}{\rho_{ds}^2 - 1} z_s \tag{4.14}$$

may be used in a similar way.

4.4.5 Type 1 error rates

Aberrance in C_1

For SNPs aberrant in only C_1 we have $\zeta_d \neq 0$, $\zeta_c \neq 0$, $\zeta_m \neq 0$, and $\zeta_r = \zeta_s = 0$. R_A, R_B, R_C can be considered as functions of ζ_d. As $\zeta_d \to 0$, $R_A, R_B, R_C \to P_0$ (equation 4.2). As $\zeta_d \to \pm\infty$, $R_A \to \frac{\beta}{2}$, $R_B = \frac{\beta^*}{2}$ and $R_C = \frac{\beta^\perp}{2}$. For positive ζ_d both R_A and R_B are increasing (and both are symmetric in ζ_d) so $R_A < \frac{\beta}{2}$, $R_B < \frac{\beta^*}{2}$, $R_C < \frac{\beta^\perp}{2}$ for all ζ_d.

Since $\beta^\perp < \beta^* < \beta$ (often substantially), methods B and C are generally better at rejecting $H_0^\cap$ for such SNPs. In the simplified case where $z_\gamma = 1$, $R_A \geq R_B$ universally (appendix B, section B.1.3. This typically holds for all z_γ, except for small deviations in pathological cases.

In general, I consider aberrance which is only still present after any strata or covariates have been accounted for in the computation of z scores. If strata or covariates remove the effective aberrance between groups, the type-1 error rate is equivalent to that under $H_0^{\cap}$.

Aberrance in C_1'

For SNPs aberrant in C_1', we have $\zeta_d = 0$, $\zeta_c = 0$, $\zeta_r \neq 0$, $\zeta_s \neq 0$ and $\zeta_m \neq 0$.

Again, $R_A, R_B, R_C \to P_0$ as $\zeta_r \to 0$. As $\zeta_r \to \pm\infty$, $R_A, R_B, R_C \to \frac{\alpha}{2}$, and both are bounded by $\frac{\alpha}{2}$. Although R_B and R_C are typically higher than R_A in this case, since both have the same (typically conservative) upper bound, this is not typically a large sacrifice in type 1 error.

In the simplified case where $\gamma = 1$, an approximate upper bound on $R_B - R_A$ is given by (appendix B, section B.1.4)

$$\frac{\alpha}{2\sqrt{2\pi}}\left(\frac{k}{\sqrt{1-\rho^2}} - 1\right) z_\beta \ll \frac{\alpha}{2} \tag{4.15}$$

where

$$k = \frac{\zeta_s}{\zeta_r} \approx \sqrt{\frac{(n_0 + n_0')(n_0' + n_1')}{n_0'(n_0 + n_0' + n_1')}} \tag{4.16}$$

In practice, there is typically a similarly small difference between R_C, R_B and R_A in the general case.

Aberrance in C_0'

For SNPs aberrant in C_0', $\zeta_d = 0$, $\zeta_r \neq 0$, $\zeta_c \neq 0$, $\zeta_s \neq 0$ and $\zeta_m \neq 0$. As for SNPs with aberrance in C_1', $R_A, R_B, R_C \to P_0$ as $\zeta_r \to 0$ and as $\zeta_r \to \pm\infty$, $R_A, R_B \to \frac{\alpha}{2}$, both bounded above by $\frac{\alpha}{2}$. R_C, however, tends to 1 as $\zeta_d \to \infty$.

In method B the cohort C_0 has a correcting effect on the replication study, meaning $|\zeta_s| < |\zeta_r|$ and $R_B < R_A$.

For the simplified case where $\gamma = 1$, a similar bound to 4.15 holds for the difference $R_A - R_B$ (note signs are reversed) with

$$k' = \frac{\zeta_s}{\zeta_r} \approx \sqrt{\frac{n_0'(n_0' + n_1')}{(n_0 + n_0')(n_0 + n_0' + n_1')}} \tag{4.17}$$

in the place of k. The improvement in type-1 error rate for a SNP with aberrance in C_0' is generally larger than the loss with the same aberrance in C_1' (see methods), meaning that if

aberrances are of similar prevalence and size in C_1' and C_0', method B will typically have a lower type-1 error rate than method A.

Aberrance in C_0

Aberrance in C_0 represents a serious problem in case-control study comparison. False-positive rates are generally worse under method B, and tend to 1 as $E(z) \to \infty$. If aberrances of this type are expected to be very frequent, this may preclude use of methods B or C.

However, aberrances of this type may be best detected retrospectively by examining aberrances between control groups at SNPs declared 'hits'. This procedure is already a necessary quality-control procedure in method A [The Wellcome Trust Case Control Consortium, 2007, Anderson et al., 2010], as method A does not provide any control over differences between C_0 and C_0'. The number of SNPs reaching significance in the two-stage procedure is usually small enough that this examination is readily tractable.

Aberrance in two or more cohorts

If SNPs are aberrant in both C_1 and C_1', or in both C_0 and C_0', the effect on R_A and R_B is similar. If both cohorts are aberrant in the same direction, there is no way to differentiate the SNP from a genuine association on the basis of the genotype data alone. If cohorts are aberrant in different directions, then in both methods, the type-1 error rate is lower than for a null SNP with no aberration or aberration in only one cohort, as effect sizes for the discovery and replication cohorts are biased in opposite directions. The same typically holds if C_0' and C_1, or C_0 and C_1', are biased in the same direction.

If C_0' and C_1' or C_0 and C_1 are both biased in the same direction, R_A is generally lower than R_B, as $\zeta_s \neq 0$. Both R_A and R_B are bounded by $\frac{\alpha}{2}$ in this case. In addition, a systematic bias in both replication groups (or both discovery groups) is likely to be due to a known confounder, the effect of which can be removed by performing a stratified test (as is typically good practice when confounders are known). Aberrance in opposite directions leads to $R_B > R_A$ in the first case, and a scenario similar to aberrance in C_0 in the second case.

Aberrance in three or more cohorts corresponds to a chaotic scenario in which neither methods A,B, or C will reliably provide FPR control. Aberrance of this extent is typically detectable and removable using quality control procedures.

Chapter 5

Characterising disease heterogeneity

5.1 Introduction

Analysis of genetic data in human disease typically uses a binary disease model of cases and controls. However, many common human diseases show extensive clinical and phenotypic diversity which may represent multiple causative pathophysiological processes. Because therapeutic approaches often target disease-causative pathways, understanding this phenotypic complexity is valuable for further development of treatment, and the progression towards personalised medicine. Indeed, identification of patient subgroups characterised by different clinical features can aid directed therapy [Li et al., 2015] and accounting for phenotypic substructures can improve ability to detect causative variants by refining phenotypes into subgroups in which causative variants have larger effect sizes [Morris et al., 2009].

Such subgroups may arise from environmental effects, reflect population variation in non-disease related anatomy or physiology, correspond to partitions of the population in which disease heritability differs, or represent different causative pathological processes. In this chapter, I present a method which tests whether there exist a subset of disease-associated SNPs which have different effect sizes in case subgroups, determining whether heterogeneity corresponds to differential genetic pathology.

The test is for a stronger assertion than the question of whether subgroups of a disease group exhibit any genetic differences at all, as these may be entirely disease-independent: for example, although there will be systematic genetic differences between Asian and European patient cohorts with type 1 diabetes (T1D), these differences will not generally relate to the pathogenesis of disease.

Rather than attempting to analyse SNPs individually for differences between subgroups, a task for which GWAS are typically underpowered, I model allelic differences across all

SNPs using mixture multivariate normal models. This can give insight into the 'geometric' structure of the genetic basis for disease. Given evidence that there exists some subset of SNPs that both differentiate controls and cases and differentiate subgroups, I present a method to reassess test statistics to search for single-SNP effects.

5.2 Results

5.2.1 Summary of proposed method

I jointly consider allelic differences between the combined case group and controls, and allelic differences between case subgroups independent of controls. Specifically, I establish whether the data support a hypothesis (H_1) that a subset of SNPs associated with case-control status have different underlying effect sizes (and hence underlying allele frequencies) in case subgroups. This assumption has been used previously for genetic discovery [Plagnol et al., 2011].

H_1 encompasses several potential underlying mechanisms of heterogeneity. A set of SNPs may be associated with one case subgroup but not the other; the same set of SNPs may have different relative effect sizes in subgroups, or heritability may differ between subgroups. These scenarios are discussed in appendix C, section C.1.

The overall protocol is to fit two bivariate Gaussian mixture models, corresponding to null and alternative hypotheses, to summary statistics (Z scores) derived from SNP data. I assume a group of controls and two non-intersecting case subgroups, and jointly consider allelic differences between the combined case group and controls, and allelic differences between case subgroups independent of controls (figure 5.1). Heterogeneity in cases can also be characterised by a quantitative trait, rather than explicit subgroups.

For a given SNP I denote by μ_1, μ_2, μ_{12} and μ_c the population minor allele frequencies for each of the two case subgroups, the whole case group and the control group respectively, and P_d, P_a GWAS p-values for comparisons of allelic frequency between case subgroups and between cases and controls, under the null hypotheses $\mu_1 = \mu_2$ and $\mu_{12} = \mu_c$ respectively (or similarly for quantitative heterogeneity). I then derive absolute Z scores $|Z_d|$ and $|Z_a|$ from these p-values (see figure 5.1). The values $|Z_d|, |Z_a|$ are considered as absolute values of observations of random variables (Z_d, Z_a) which are samples from a mixture of three bivariate Gaussians. Further details are given in appendix C, section C.2.

I consider SNPs to be partitioned into three categories, with each category corresponding to a different joint distribution of Z_d, Z_a:

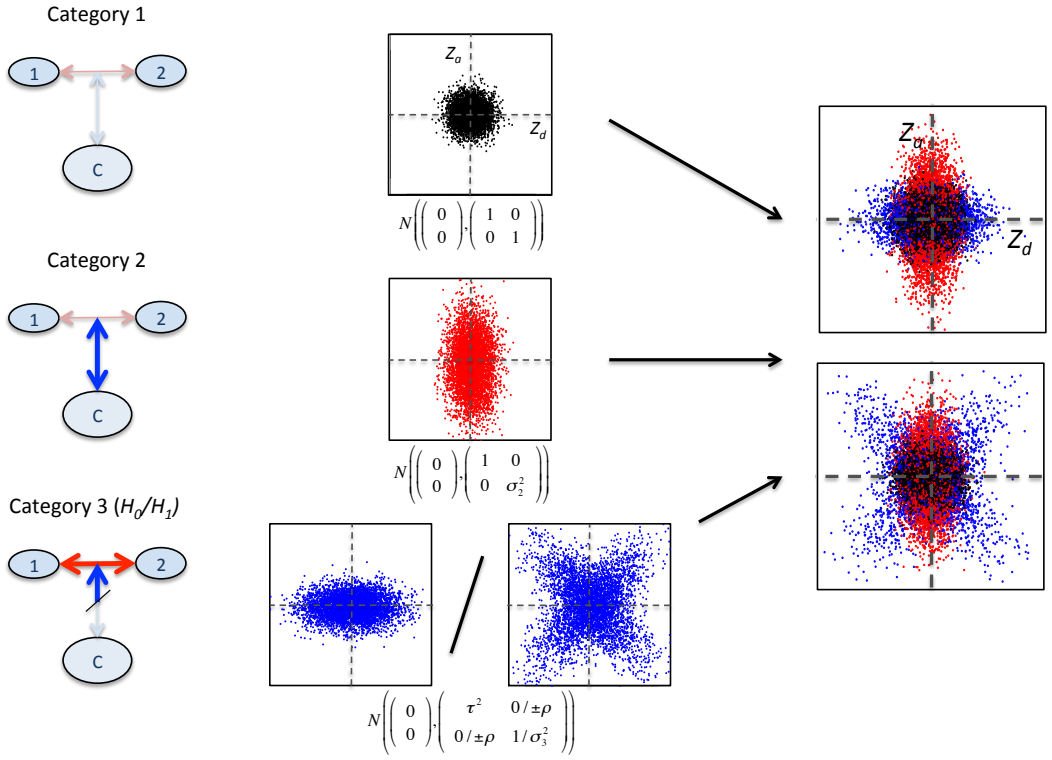


Fig. 5.1 Overview of three-categories model. Z_d and Z_a are Z scores derived from GWAS p-values for allelic differences between case subgroups (1 vs 2), and between cases and controls ($1+2$ vs C) respectively (left). Within each category of SNPs, the joint distribution of (Z_d, Z_a) has a different characteristic form. In category 1, Z scores have a unit normal distribution; in category 2, the marginal variance of Z_a can vary. The distribution of SNPs in category 3 depends on the main hypothesis. Under H_0 (that all disease-associated SNPs have the same effect size in both subgroups), only the marginal variance of Z_d may vary; under H_1 (that subgroups correspond to differential effect sizes for disease-associated SNPs), any covariance matrix is allowed. The overall SNP distribution is then a mixture of Gaussians resembling one of the rightmost panels, but with SNP category membership unobserved. Visually, my test determines whether the observed overall Z_d, Z_a distribution more closely resembles the bottom rightmost panel than the top.

1. SNPs which do not differentiate subgroups and are not associated with the phenotype as a whole ($\mu_c = \mu_1 = \mu_2$)

2. SNPs which are associated with the phenotype as a whole but which are not differentially associated with the subgroups ($\mu_c \neq \mu_{12}$; $\mu_1 = \mu_2 = \mu_{12}$)

3. SNPs which have different population allele frequencies in subgroups, and may or may not be associated with the phenotype as a whole ($\mu_1 \neq \mu_2$)

If the SNPs in category 3 are not associated with the disease as a whole (null hypothesis, H_0), we expect Z_d, Z_a to be independent and the variance of Z_a to be 1. If SNPs in category 3 are also associated with the disease as a whole (alternative hypothesis, H_1), the joint distribution of (Z_d, Z_a) will have both marginal variances greater than 1, and Z_a, Z_d may covary. My test is therefore focussed on the form of the joint distribution of (Z_d, Z_a) in category 3. Importantly, I allow that the correlation between Z_d and Z_a may be simultaneously positive at some SNPs and negative at others. This allows for a subset of SNPs to specifically alter risk of one subgroup, and another subset to alter risk for the other subgroup. To accommodate this, I only consider absolute Z scores and model the distribution of SNPs in category 3 with two mirror-image bivariate Gaussians.

Amongst SNPs with the same frequency in disease subgroups (categories 1 and 2), Z_a and Z_d are independent and the expected standard deviation of Z_d is 1. I therefore model the overall joint distribution of (Z_d, Z_a) as a Gaussian mixture in which the PDF of each observation (Z_d, Z_a) is given by

$$\begin{aligned} PDF_{Z_d,Z_a|\Theta}(d,a) =& \pi_1 N_{\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)}(d,a) && \text{(category 1)} \\ &+ \pi_2 N_{\left(\begin{smallmatrix}1 & 0\\ 0 & \sigma_2^2\end{smallmatrix}\right)}(d,a) && \text{(category 2)} \\ &+ \pi_3 \left(\frac{1}{2} N_{\left(\begin{smallmatrix}\tau^2 & \rho\\ \rho & \sigma_3^2\end{smallmatrix}\right)}(d,a) + \frac{1}{2} N_{\left(\begin{smallmatrix}\tau^2 & -\rho\\ -\rho & \sigma_3^2\end{smallmatrix}\right)}(d,a) \right) && \text{(category 3)} \end{aligned} \quad (5.1)$$

where $N_\Sigma(d,a)$ denotes the density of the bivariate normal PDF centred at $\left(\begin{smallmatrix}0\\0\end{smallmatrix}\right)$ with covariance matrix Σ at (d,a). Θ is the vector of values $(\pi_1, \pi_2, \tau, \sigma_2, \sigma_3, \rho)$. Under H_0, we have $\rho = 0$ and $\sigma_3 = 1$. The values (π_1, π_2, π_3) represent the proportion of SNPs in each category, with $\Sigma\pi_i = 1$ (see table 5.1). Patterns of (Z_d, Z_a) for different parameter values are shown in appendix D.1, table D.1.

	Model	Interpretation
π_1	H_0/H_1	Proportion of SNPs **not** associated with case/control status and **not** associated with subgroup status (category 1)
π_2	H_0/H_1	Proportion of SNPs associated with case/control status but **not** subgroup status (category 2)
π_3	H_0/H_1	Proportion of SNPs associated with subgroup status (category 3)
τ	H_0/H_1	Standard deviation of observed Z_d scores (effect sizes for subgroup status) in category 3
σ_2	H_0/H_1	Standard deviation of observed Z_a scores (effect sizes for case/control status) in category 2
σ_3	H_1 only	Standard deviation of observed Z_a scores (effect sizes for case/control status) in category 3
ρ	H_1 only	'Absolute covariance' between Z_d scores (effect sizes for subgroup status) and Z_a scores (effect sizes for case/control status) in category 3

Table 5.1 Interpretation of parameter values in the fitted model. Parameters τ, σ_2 and σ_3 are dependent on sample sizes, but can be converted to sample-size independent forms (see appendix C, section C.4.3).

I use the product of values of the above PDF for a set of observed Z_d, Z_a as an objective function ('pseudo-likelihood', PL) to estimate the values of parameters. This is not a true likelihood as observations are dependent due to linkage disequilibrium (LD), although because I minimise the degree of LD between SNPs using the LDAK method [Speed et al., 2012], the PL is similar to a true likelihood.

5.2.2 Model fitting and significance testing

Parameters π_1, π_2, π_3 $(= 1 - \pi_1 - \pi_2)$, σ_2, σ_3, τ and ρ are fit under H_1 and H_0. Under H_0, $(\rho, \sigma_3) = (0, 1)$.

The fit of the two models is compared using the log-ratio of PLs, giving an unadjusted pseudo-likelihood ratio (uPLR). A term depending only on Z_a is subtracted to minimise the influence of the Z_a score distribution, and a term $\log(\pi_1\pi_2\pi_3)$ added to ensure the model is identifiable [Chen et al., 2001]. I term the resultant test statistic the pseudo-likelihood ratio (PLR). The null distribution of the PLR is conservatively estimated by a distribution of the

form:

$$PLR|H_0 \sim \begin{cases} \gamma\chi^2_1 & \text{prob} = \kappa \\ \gamma\chi^2_2 & \text{prob} = 1-\kappa \end{cases} \tag{5.2}$$

The value γ arises from the weighting derived from the LDAK procedure causing a scale change in the observed *PLR*. The mixing parameter κ corresponds to the probability that $\rho = 0$, (approximately $\frac{1}{2}$).

I estimate γ and κ by sampling random subgroups of the case group. Such subgroups only cover the subspace of H_0 with $\tau = 1$ (no systematic allelic differences between subgroups), causing the asymptotic approximation of PLR by equation 5.2 to be poor. I thus estimate γ and κ from the distribution of a similar alternative test statistic, the cPLR (see methods section and appendix C, section C.3), which is well-behaved even when $\tau \approx 1$ and enables a conservative estimate of a p-value under H_0.

A natural next step is to search for the specific variants contributing to the PLR. An effective test statistic for testing subgroup differentiation for single SNPs is the Bayesian conditional false discovery rate (cFDR), introduced in chapter 2 [Andreassen et al., 2013, Liley and Wallace, 2015] applied to Z_d scores 'conditioned' on Z_a scores. However, this statistic alone cannot capture all the means by which the joint distribution of (Z_a, Z_d) can deviate from H_0, and we also propose three other test statistics, each with different advantages, and compare their performance (appendix C, section C.6.1).

5.2.3 Power calculations, simulations, and validation of method

I tested the method by application to a range of datasets, using simulated and re-sampled GWAS data. First, to confirm appropriate control of type 1 error rates across H_0, I simulated genotypes of case and control groups under H_0 for a set of 5×10^5 autosomal SNPs in linkage equilibrium (appendix C, section C.4). Quantiles of the empirical PLR distribution were smaller than those for the empirical cPLR distribution and the asymptotic mixture-χ^2, indicating that the test is conservative when $\tau > 1$ (estimated type 1 error rate 0.048, 95% CI 0.039-0.059) and when $\tau \approx 1$ (estimated type 1 error rate 0.033, 95% CI 0.022-0.045) as expected; see figure 5.2. The distribution of cPLR closely approximated the asymptotic mixture-χ^2 distribution across all values of τ (appendix C, figure C.5).

I then established the suitability of the test when SNPs are in LD and when there exist genetic differences between subgroups that are independent of disease status overall. First, I used a dataset of controls and autoimmune thyroid disease (ATD) cases and repeatedly

choose subgroups such that several SNPs had large allelic differences between subgroups. I found good FDR control at all cutoffs (appendix D.2, figure C.6) and the overall type 1 error rate at $\alpha = 0.05$ was 0.041 (95% CI 0.034-0.050). Second, I analysed a dataset of T1D cases with subgroups defined by geographical origin. Within the UK, there is clear genetic diversity associated with region [Leslie et al., 2015]. As expected, Z_d scores for geographic subgroups showed inflation compared to for random subgroups (appendix D.2, figure D.1). None of the derived test statistics reached significance at a Bonferroni-corrected $p < 0.05$ threshold (min. corrected p value > 0.8, appendix D.2, figre D.2).

To examine the power of my method, I used published GWAS data from the Wellcome Trust Case Control Consortium [The Wellcome Trust Case Control Consortium, 2007] comprising 1994 cases of Type 1 diabetes (T1D), 1903 cases of rheumatoid arthritis (RA), 1922 cases of type 2 diabetes (T2D) and 2953 common controls. I established that the test could differentiate between any pair of diseases, considered as subgroups of a general disease case group (all $< 1 \times 10^{-8}$, table 5.2).

T1D and RA have overlap in genetic basis [The Wellcome Trust Case Control Consortium, 2007, Fortune et al., 2015, Liley and Wallace, 2015], as well as non-overlapping associated regions. T1D and T2D have less overlap [Fortune et al., 2015] and T2D and RA less still. This was reflected in the fitted values (table 5.2, figure 5.3). The fitted values parametrizing category 2 in the full model for T1D/RA (π_2, σ_2) were consistent with a subset of SNPs associated with case/control status (T1D+RA vs control) but not differentiating T1D/RA. By contrast, the parametrization of category 2 for T1D/T2D and T2D/RA had marginal variance σ_2 approximately 1, suggesting that a subset of SNPs associated with case/control status but not with 'subgroup' status did not exist in these cases. The rejection of H_0 for the comparisons entails the existence of a set of SNPs associated both with case/control and subgroup status. The H_0 model does not allow such a set of SNPs, forcing the parametrisation of Z_d, Z_a scores for such SNPs to be 'squashed' into a category shape permitted under H_0, with one marginal variance being 1: either category 2 (as happens in T2D/RA since $\pi_2|H_0 \approx \pi_3|H_1$, $\sigma_2|H_0 \approx \sigma_3|H_1$ in T2D/RA) or category 3 (as in T1D/T2D, where $\pi_3|H_0 \approx \pi_3|H_1$, $\tau|H_0 \approx \tau|H_1$).

To determine the power of the test more generally, I showed that power depends on the number of SNPs in category 3 and on the underlying parameters of the true model, depending on the number of samples through the fitted model parameters (appendix C, section C.4.3). I therefore estimated the power of the test for varying numbers of SNPs in category 3 and for varying values of the parameters σ_3, τ, and ρ. (Figure 5.4; appendix D.2, figure D.3). As expected, power increases with an increasing number of SNPs in category 3, reflecting the

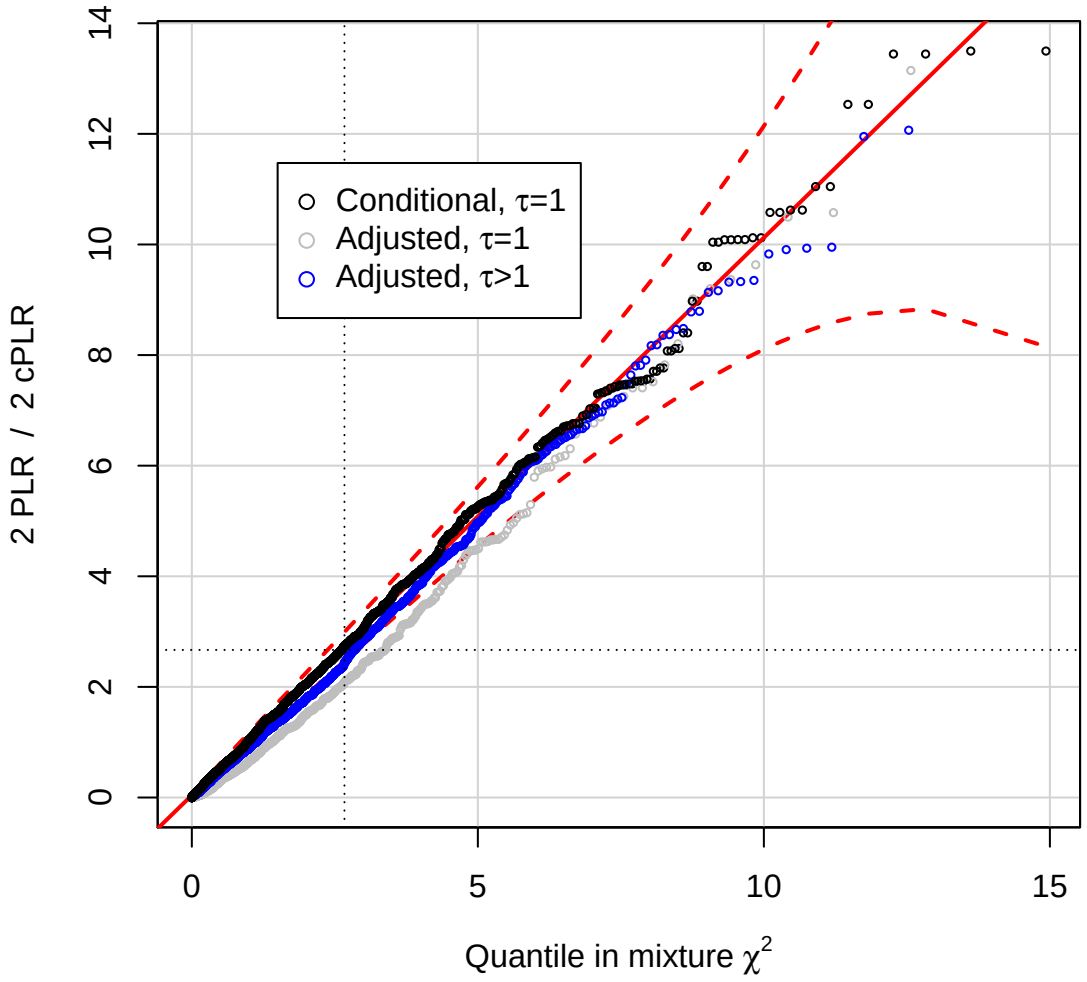


Fig. 5.2 QQ plot from simulations demonstrating type 1 error rate control of PLR test. PLR values for test subgroups under H_0 with either $\tau = 1$ (random subgroups; grey) or $\tau > 1$ (genetic difference between subgroups, but independent of main phenotype; blue) with cPLR values for random subgroups (black) and against proposed asymptotic distribution under simulation ($\frac{1}{2}\left(\chi^2_1 + \chi^2_2\right)$; solid red line; 99% confidence limits dashed red line). The distribution of cPLR for random subgroups generally majorises the distribution of PLR (that is, $Pr(cPLR > x|H_0) > Pr(PLR > x|H_0)$ for all x), meaning the cPLR-based test is conservative relative to testing against an empirical distribution of PLR values from random subgroups. Further details are shown in appendix C, sections C.3 and C.4.

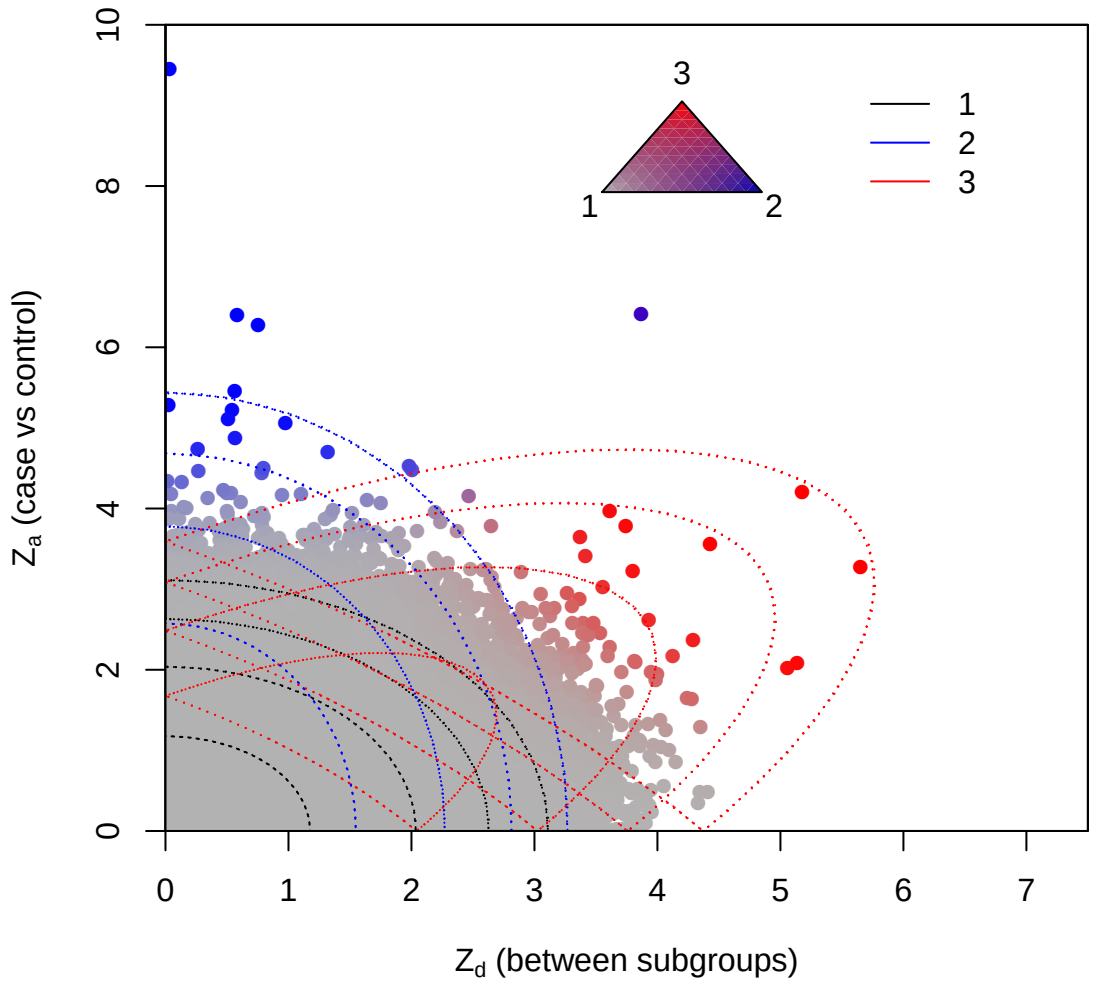


Fig. 5.3 Observed absolute Z_a and Z_d for T1D/RA. Colourings correspond to posterior probability of category membership under full model (see triangle): grey - category 1, blue - category 2, red -category 3. Contours of the component Gaussians in the fitted full model are shown by dotted lines.

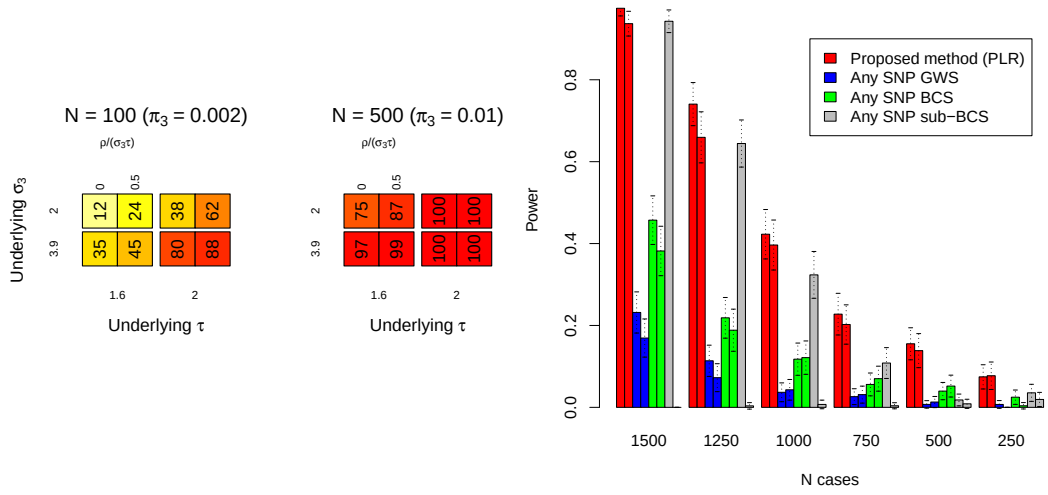


Fig. 5.4 Power of PLR to reject H_0 (genetic homogeneity between subgroups) depends on the number of SNPs in category 3 and the underlying values of model parameters σ_2, σ_3, τ, ρ. Dependence on number of case/control samples arises through the magnitudes of σ_3 and τ (appendix C, section C.4.3). Leftmost figure shows power estimates for various values of π_3, σ_3, τ, ρ. Value N is the approximate number of SNPs in category 3, ($\propto \pi_3$). Each simulation was on 5×10^4 simulated autosomal SNPs in linkage equilibrium. Value $\rho/(\sigma_3\tau)$ is the absolute correlation between Z_d and Z_a in category 3. Also see appendix D.2, figure D.3. Rightmost figure shows power of PLR to detect differences in genetic basis of T1D and RA subgroups of a combined autoimmune dataset, down-sampling to varying numbers of cases (X axis). PLR is compared with: power to find ≥ 1 SNP with Z_d score reaching genome-wide significance (GWS, blue; $p \leq 5 \times 10^{-8}$) or Bonferroni-corrected significance (BCS, green; $p \leq 0.05/(\text{total \# of SNPs})$); and power to detect any SNP with Z_a score reaching genome-wide significance and Z_d score reaching Bonferroni-corrected significance (sub-BCS, grey; $p \leq 0.05/(\text{total \# of SNPs with } Z_a \text{ reaching GWS})$). Error bars show 95% CIs. Circles/solid lines for each colour show power for all SNPs, triangles/dashed lines for all SNPs except rs17696736. Power for sub-BCS drops dramatically but power for PLR is not markedly affected, indicating relative robustness of PLR to single-SNP effects.

Table 5.2 Fitted parameter values for models of T1D/RA, T1D/T2D, T2D/RA, and GD/HT. H_1 is the null hypothesis (under which $\sigma_3 = 1$, $\rho = 0$) that SNPs differentiating the subgroups are not associated with the overall phenotype; H_1 is the alternative (full model). p values for pseudo-likelihood ratio tests are also shown.

Subgroups		π_1	π_2	π_3	σ_2	σ_3	τ	ρ
T1D/RA	H_1	0.997	5.69×10^{-4}	2.06×10^{-3}	2.76	1.39	1.74	1.815
	H_0	0.997	6.26×10^{-4}	2.48×10^{-3}	2.71	-	1.67	-
T1D/T2D	H_1	0.573	0.426	9.63×10^{-4}	1.00	2.03	2.25	1.68
	H_0	0.578	0.421	8.91×10^{-4}	1.00	-	2.21	-
T2D/RA	H_1	0.573	0.426	8.71×10^{-4}	1.00	2.23	1.75	1.69
	H_0	0.91	8.05×10^{-4}	0.0892	2.25	-	0.97	-
GD/HT	H_1	0.506	0.487	0.007	1.12	2.90	1.65	2.61
	H_0	0.493	0.079	0.428	1.68	-	1.03	-

Subgroups	p-val
T1D/RA	3.2×10^{-12}
T1D/T2D	1.6×10^{-9}
T2D/RA	5.1×10^{-9}
GD/HT	2.2×10^{-15}

proportion of SNPs which differentiate case subgroups and are associated with the phenotype as a whole. Power also increases with increasing τ, σ_3, and absolute correlation ($\rho/(\sigma_3\tau)$) as high values enable better distinction of SNPs in the second and third categories.

I explored the dependence of power on sample size by sub-sampling the WTCCC data for RA and T1D (figure 5.4) and compared the power of the PLR with the power to find any single SNP which differentiated the two diseases in several ways (see figure legend). Although the power of the PLR-based test was limited at reduced sample sizes, it remained consistently higher than the power to detect any single SNP which differentiated the two diseases. I then repeated the analysis removing the known T1D- and RA- associated SNP rs17696736. The power to detect a SNP with significant Z_d score (Bonferroni-corrected) amongst SNPs with GW-significant Z_a score dropped dramatically, though the power of PLR was only slightly reduced. This illustrated the robustness of the PLR test to inclusion

or removal of single SNPs with large effect sizes, a property not shared by single-SNP approaches.

Estimating power requires an estimate of the underlying values of several parameters: the expected total number of SNPs in the pruned dataset with different population MAF in case subgroups, and the distribution of odds-ratios such SNPs between subgroups and between cases/controls. With sparse genome-wide cover, such as that in the WTCCC study, > 1250 cases per subgroup are necessary for 90% power (discounting MHC region). If SNPs with greater coverage for the disease of interest are used (such as the ImmunoChip for autoimmune diseases) values of π_3, σ_3 and τ are correspondingly higher, and around 500-700 cases per subgroup may be sufficient.

5.2.4 Application to autoimmune thyroid disease and type 1 diabetes

Autoimmune thyroid disease (ATD) takes two major forms: Graves' disease (GD; hyperthyroidism) and Hashimoto's Thyroiditis (HT; hypothyroidism). Differential genetics of these conditions have been investigated. Detection of individual variants with different effect sizes in GD and HT is limited by sample size (particularly HT); however, the *TSHR* region shows evidence of differential effect [Cooper et al., 2012]. T1D is relatively clinically homogeneous with no major recognised subtypes, although heterogeneity arises between patients in levels of disease-associated autoantibodies, and disease course differs with age at diagnosis [Plagnol et al., 2011]. I analysed both of these diseases.

For ATD, I was able to confidently detect evidence for differential genetic bases for GD and HT ($p = 2.2 \times 10^{-15}$). Fitted values are shown in table 5.2. The distribution of cPLR statistics from random subgroups agreed well with the proposed mixture χ^2 (appendix D.2, figure D.5b).

For T1D, I considered four subgroupings defined by plasma levels of the T1D-associated autoantibodies thyroid peroxidase antibody (TPO-Ab, n=5780), insulinoma-associated antigen 2 antibody (IA2-Ab, n=3197), glutamate decarboxylase antibody (GAD-Ab, n=3208) and gastric parietal cell antibodies (PCA-Ab, n=2240). A previous GWAS study on autoantibody positivity in T1D identified only two non-MHC loci at genome-wide significance: 1q23/*FCRL3* with IA2-Ab and 9q34/*ABO* with PCA-Ab [Plagnol et al., 2011].

I tested each of the subgroupings retaining and excluding the MHC region. Fitted values for models with and without MHC are shown in appendix D.1, figure D.2, and plots of Z_a and Z_d scores are shown in appendix D.2, figure D.6. Retaining the MHC region, I was able to confidently reject H_0 for subgroupings based on TPO-Ab, IA-2Ab and GAD-Ab

(all p-values $< 1.0 \times 10^{-20}$). Although there was evidence that SNPs in the dataset were associated with PCA-Ab level ($\tau \approx 2.5$, null model), the improvement in fit in the full model was not significant, and I conclude that such SNPs determining PCA-Ab status are not in general T1D-associated. This can be seen by in the plot of Z_a against Z_d (appendix D.2, figure D.6) where SNPs with high Z_d values do not have higher than expected Z_a values.

With MHC removed, the subgrouping on TPO-Ab was significantly better-fit by the full model ($p = 1.5 \times 10^{-4}$). There was weaker evidence to reject H_0 for GAD-Ab ($p = 0.002$) and IA2-Ab ($p = 0.008$) (Bonferroni-corrected threshold at $\alpha < 0.05$: 0.006). Fitted values of τ in both the full and null models for GAD-Ab were ≈ 1, indicating absence of evidence for a category of non-MHC T1D-associated SNPs additionally associated with GAD-Ab positivity. Collectively, this indicates that differential genetic basis for T1D with GAD-Ab and IA2-Ab positivity is driven principally by the MHC region, and although PCA-Ab status is partially genetically determined, the set of causative variants is independent of T1D causative pathways.

The variation in genetic architecture of T1D with age is not fully understood, but previous studies have suggested larger observed effects at known loci in patients diagnosed at a younger age [Hyttinen et al., 2003, Howson et al., 2009, Howson et al., 2011, Howson et al., 2012]. I investigated whether these differences were indicative of widespread differences in variant effect sizes with age-at-diagnosis, possibly due to differential heritability (see appendix C, section C.1). I applied the method to T1D dataset with Z_d defined by age at diagnosis (quantitative trait). Fitted values are shown in appendix D.1, table D.3 and Z_a and Z_d scores in appendix D.2, figure D.7. The hypothesis H_0 could be rejected confidently when retaining or removing the MHC region (p values $< 1.0 \times 10^{-20}$ and 0.007 respectively). Signed Z_d and Z_a scores for age at diagnosis showed a visible negative correlation ($p = 0.002$) amongst Z_d and Z_a scores for disease-associated SNPs (r_g method 2, figure 5.5). This is consistent with a higher genetic liability with lower age at diagnosis.

5.2.5 Assessment of individual SNPs

Many SNPs which discriminated subgroups were in known disease-associated regions (appendix D.1, tables D.6, D.9, and D.12). In several cases, the new method identified disease-associated SNPs which have reached genome-wide significance in subsequent larger studies but for which the Z_a score in the WTCCC study was not near significance. For example, the SNP rs3811019, in the *PTPN22* region, was identified as likely to discriminate

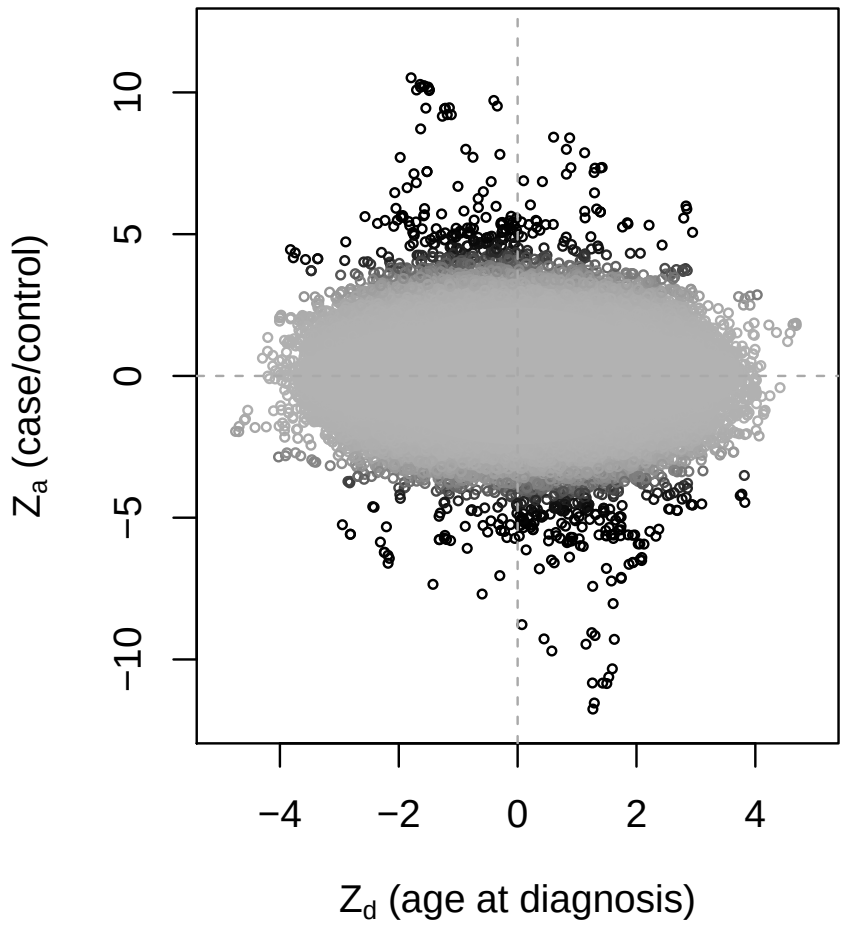


Fig. 5.5 Z_a and Z_d scores for age at diagnosis in T1D, excluding MHC region. Colour corresponds to posterior probability of category 2 membership in null model (since categories in full model are assigned on the basis of correlation), with black representing a high probability. Z_d and Z_a are negatively correlated ($p = 8.7 \times 10^{-5}$ with MHC included, $p = 0.002$ with MHC removed) after accounting for LD using LDAK weights, and weighting by posterior probability of category 2 membership in the null model, to prioritise SNPs further from the origin

T1D and T2D ($p = 3.046 \times 10^{-6}$; appendix D.1, table D.9), despite a p value of 3×10^{-4} for joint T1D/T2D association.

For GD and HT, SNPs near the known ATD-associated loci *PTPN22* (rs7554023), *CTLA4* (rs58716662), and *CEP128* (rs55957493) were identified as likely to be contributing to the difference (see appendix D.1, table D.15). The SNPs rs34244025 and rs34775390 are not known to be ATD-associated, but are in known loci for inflammatory bowel disease and ankylosing spondylitis, and this analysis suggest they may differentiate GD and HT (FDR 0.003).

I searched for non-MHC SNPs with differential effect sizes with TPOA positivity in T1D, the subgrouping of T1D for which we could most confidently reject H_0. Previous work [Plagnol et al., 2011] identified several loci potentially associated with TPO-Ab positivity by restricting attention to known T1D loci, enabling use of a larger dataset than was available to us. I list the top ten SNPs for each summary statistic for TPO-Ab positivity in appendix D.1, table D.18. Subgroup-differentiating SNPs included several near known T1D loci: *CTLA4* (rs7596727), *BACH2* (rs11755527), *RASGRP1* (rs16967120) and *UBASH3A* (rs2839511) [Barrett et al., 2009]. These loci agreed with those found by Plagnol et al [Plagnol et al., 2011], but my analysis used only available genotype data, without external information on confirmed T1D loci. I was not able to replicate the same p-values due to reduced sample numbers.

Finally, I analysed non-MHC SNPs with varying effect sizes with age at diagnosis in T1D (appendix D.1, table D.21). This implicated SNPs in or near *CTLA4* (rs2352551), *IL2RA* (rs706781), and *IKZF3* (rs11078927).

5.3 Discussion

The problem I address is part of a wider aim of adapting GWAS to complex disease phenotypes. As the body of GWAS data grows the analysis of between-disease similarity and within-disease heterogeneity has led to substantial insight into shared and distinct disease pathology [Andreassen et al., 2013, Liley and Wallace, 2015, Morris et al., 2009, Traylor et al., 2013, Wen and Lu, 2013]. I seek in this paper to use genomic data to infer whether such disease subtypes exist. The problem is related to the question of whether two different diseases share any genetic basis [Bulik-Sullivan et al., 2015] but differs in that the implicit null hypothesis relates to genetic homogeneity between subgroups rather than genetic independence of separate diseases.

The test strictly assesses whether a set of SNPs have different effect sizes in case subgroups. I interpret this as 'differential causative pathology', which encompasses several disease mechanisms, discussed in appendix C, section C.1. In some cases, if subgroups are defined on the basis of the presence or absence of a known disease risk factor, the heritability of the disease will differ between subgroups, with corresponding changes in variant effect sizes.

I use 'absolute covariance' ρ preferentially (see appendix D.1, table D.1) because I expect that Z_a and Z_d will frequently co-vary positively and negatively at different SNPs in the same analysis; for instance, if some variants are deleterious only for subgroup 1 and others only for subgroup 2. A potential advantage of the symmetric model is the potential to generate Z_d scores from ANOVA-style tests for genetic homogeneity between three or more subgroups, in which case reconstructed Z scores would be directionless. This situation is similar to the analysis described in section 6, although I use alternative methods in that case.

Aetiologically and genetically heterogeneous subgroups within a case group correspond to substructures in the genotype matrix. As discussed in chapter 1, section 1.4.5, information about such substructures is lost in a standard GWAS, which only uses the column-sums (MAFs) of the matrix (linear-order information). Data-driven selection of appropriate case subgroups and corresponding analyses of these subgroups can use more of the remaining quadratic-order information the matrix contains. Indeed a 'two-dimensional' GWAS approach (using Z_a and Z_d) instead of a standard GWAS (using only Z_a) may improve SNP discovery, as I found for *PTPN22* in RA/T2D. However, this can only be the case if the subgroups correspond to different variant effect sizes; for other subgroupings, a two-dimensional GWAS will only add noise.

While it seems appealing to use this method to search for some 'optimal' partition of patients, I focus on testing subgroupings derived from independent clinical or phenotypic data. Firstly, it is difficult to characterise subgroupings as 'better' or 'worse', and no one parameter can parametrise the degree to which two subgroups differ; parameters π_3, τ, and ρ all contribute, and attempts to test the hypothesis using a single measure such as genetic correlation have serious shortcomings (appendix C, section C.5). Secondly, even if subgroups could meaningfully be ranked, the search space of potential subgroupings of a case group is prohibitively large (2^N for N cases), making exhaustive searches difficult. This is explored further in chapter 7, section 7.1.5.

I demonstrated that effect sizes of T1D-causative SNPs differ with age at disease diagnosis. The strong negative correlation observed (figure 5.5) was consistent with an increased total genetic liability in samples with earlier age of diagnosis, a finding supported by can-

didate gene studies [Howson et al., 2009, Howson et al., 2011, Howson et al., 2012] and epidemiological data [Hyttinen et al., 2003]. Such a pattern arises naturally from a liability threshold model where total liability depends additively on both genetic effects and environmental influences which accumulate with age (appendix C, section C.1).

The method necessarily dichotomises the multitude of mechanisms of heterogeneity, although there are many diverse forms (appendix D.1, table D.1; appendix C, section C.1). There is potential to further dissect the mechanisms of disease heterogeneity by incorporating estimations of genetic correlation [Bulik-Sullivan et al., 2015] or assessing evidence for liability threshold models [Chatterjee and Carroll, 2005]. Similar mixture-Gaussian approaches may also be adaptable to this purpose, by assessing other families of effect size distributions.

The methods in this chapter add to the current body of knowledge by extracting additional information from a disease dataset over a standard GWAS analysis, and help to determine if further analysis of disease pathogenesis in subgroups is justified. My approach is analogous to the intuitive method of searching for between-subgroup differences in SNPs with known disease associations [Plagnol et al., 2011] but does not restrict attention to strong disease associations, enabling use of information from disease-associated SNPs which do not reach significance. My parametrisation of effect size distributions allows insight into the structure of the genetic basis of the disease and potential subtypes, improving understanding of genotype-phenotype relationships.

5.4 Methods

5.4.1 Joint distribution of variables Z_a, Z_d

I assume that SNPs may be divided into three categories, as described in the results section (figure 5.1). Under these assumptions, Z_a and Z_d scores have the joint PDF given by equation 5.1.

Define Θ as the vector of values $(\pi_1, \pi_2, \pi_3, \tau, \sigma_2, \sigma_3, \rho)$. Z scores Z_a and Z_d are reconstructed from GWAS p-values for SNP associations. In practice, since the model is symmetric, I only require absolute Z scores, without considering effect direction.

If a set of SNPs have normally-distributed log-odds ratios, for which Z scores are generated, then if α is the 97.5% quantile of the odds-ratio distribution, and we have sample sizes n_1, n_2, the expected observed standard deviation of Z scores (corresponding to, σ_2, σ_3,

and τ) is given by

$$E\{SD(Z)\} = \sqrt{1 + \frac{\log(\alpha)^2 n_1 n_2}{12(n_1+n_2)}} \tag{5.3}$$

(appendix C, section C.4.3).

5.4.2 Definition and distribution of PLR statistics

For a set of observed Z scores (Z_a, Z_d) I define the joint unadjusted pseudo-likelihood $PL_{da}(Z|\Theta)$ as

$$\log\{PL_{da}(Z_d, Z_a|\Theta)\} = \sum_{Z_d^{(i)} \in Z_d, Z_a^{(i)} \in Z_a} w_i PDF_{Z_d, Z_a|\Theta}(Z_d^{(i)}, Z_a^{(i)}) + C\log(\pi_1\pi_2\pi_3) \tag{5.4}$$

where the term $C\log(\pi_1\pi_2\pi_3)$ is included to ensure identifiability of the model [Chen et al., 2001] and weights w_i are included to adjust for LD (see below).

I now set

$$\begin{aligned} \widehat{\theta_1} &= \arg\max_{\theta \in H_1} PL_{da}(Z_d, Z_a|\theta) \\ \widehat{\theta_0} &= \arg\max_{\theta \in H_0} PL_{da}(Z_d, Z_a|\theta) \\ uPLR(Z) &= \log\left(\frac{PL_{da}(Z|\widehat{\theta_1})}{PL_{da}(Z|\widehat{\theta_0})}\right) \end{aligned} \tag{5.5}$$

recalling that H_0 is the subspace of the parameter space H_1 satisfying $\sigma_3 = 1$ and $\rho = 0$.

If data observations are independent, *uPLR* reduces to a likelihood ratio. Under H_0, the asymptotic distribution of *uPLR* is then

$$uPLR \sim \frac{1}{2}\begin{cases} \chi_1^2 & p = 1/2 \\ \chi_2^2 & p = 1/2 \end{cases} \tag{5.6}$$

according to Wilk's theorem extended to the case where the null value of a parameter lies on the boundary of H_1 (since $\rho = 0$ under H_0) [Self and Liang, 1987].

The empirical distribution of *uPLR* may depart substantially from this asymptotic distribution if $\tau \approx 1$. Indeed, if X has the above asymptotic distribution, we may have $Pr(uPLR > x) \gg Pr(X > x)$ when $\tau \approx 1$ (see appendix C, sections C.3.2, C.3.3). In the full model, the marginal distribution of Z_a has more degrees of freedom (four; π_1, π_2, σ_2, σ_3) than it does under the null model (two; π_2, σ_2; as $\sigma_3 \equiv 1$). This can mean that certain distributions

of Z_a can drive high values of *uPLR* independent of the values of Z_d (appendix C, figure C.4, section C.4), which is unwanted as the values Z_a reflect only case/control association and carry no information about case subgroups. If observed *uPLR*s from random subgroups (for which $\tau = 1$ by definition) are used to approximate the null *uPLR* distribution, and a specific *uPLR* value compared against this approximated null, there would be minimal power when $\tau \gg 1$.

This effect can be managed by subtracting a correcting factor based on the pseudo-likelihood of Z_a alone, which reflects the contribution of Z_a values to the uPLR. We define

$$PL_a(Z_a|\Theta) = \prod_{Z_a^{(i)} \in Z_a} \left(\pi_1 N_{0,1}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2}(Z_a^{(i)}) + \pi_3 N_{0,\sigma_3^2}(Z_a^{(i)}) \right) \tag{5.7}$$

that is, the marginal likelihood of Z_a. Given $\widehat{\theta_1}$, $\widehat{\theta_0}$ as defined above, we define

$$f(Z_a|\hat{\theta}_1, \hat{\theta}_0) = \min \left(\log \frac{PL_a(Z_a|\widehat{\theta_1})}{PL_a(Z_a|\widehat{\theta_0})}, 0 \right) \tag{5.8}$$

We now define the PLR as

$$PLR = uPLR - f(Z_a|\hat{\theta}_1, \hat{\theta}_0) \tag{5.9}$$

Estimates $\hat{\pi}_2$ and $\hat{\sigma}_2$ of π_2 and σ_2 under H_0 can be made by fitting a bivariate distribution to the marginal observed distribution of Z_a (see section 5.4.4). Given these estimates, I define the similar test statistic *cPLR*

$$\begin{aligned} cPL(Z_d|Z_a, \theta) &= \frac{PL_{da}(Z_a, Z_d|\theta)}{PL_a(Z_a|\theta)} \\ \widehat{\theta_1^c} &= \arg\max_{\theta \in H_1 | \pi_2 = \hat{\pi}_2, \sigma_2 = \hat{\sigma}_2} cPL(Z_d|Z_a, \theta) \\ \widehat{\theta_0^c} &= \arg\max_{\Theta \in H_0 | \pi_2 = \hat{\pi}_2, \sigma_2 = \hat{\sigma}_2} cPL(Z_d|Z_a, \theta) \\ cPLR &= \log \left(\frac{cPL(Z_d|Z_a, \widehat{\theta_1^c})}{cPL(Z_d|Z_a, \widehat{\theta_0^c})} \right) \end{aligned} \tag{5.10}$$

noting that the expression $\frac{PL_{da}(Z_a,Z_d|\theta)}{PL_a(Z_a|\theta)}$ can be considered as a likelihood conditioned on the observed values of Z_a. Now

$$\begin{aligned} PLR &= \log\left(\frac{PL_{da}(Z_d,Z_a|\widehat{\theta_1})}{PL_{da}(Z_d,Z_a|\widehat{\theta_0})}\right) - \log\left(\frac{PL_a(Z_a|\widehat{\theta_1})}{PL_a(Z_d|\widehat{\theta_0})}\right) \\ &= \log\left(\frac{cPL(Z_d|Z_a,\widehat{\theta_1})}{cPL(Z_d|Z_a,\widehat{\theta_0})}\right) \end{aligned} \tag{5.11}$$

The empirical distribution of *cPLR* has approximately the same form when $\tau = 1$ or when $\tau \gg 1$ (appendix C, section C.4). The distribution is asymptotically the same as the empirical distribution of *PLR* when $\tau \gg 1$. However, when $\tau \approx 1$ (ie, for random subgroups), *cPLR* values are generally larger than *PLR* values (appendix C, section C.4), so in general $Pr(PLR > x) < Pr(cPLR > x)$. For an observed PLR value X, the approximate p-value $Pr(cPLR > X|H_0)$ is thus a conservative estimate (overestimate) of $Pr(PLR > X|H_0)$. From the above, however, the approximate p-value $Pr(cPLR > X|H_0)$ is more powerful than testing X against the observed distribution of *PLR* for random subgroups.

5.4.3 Allowance for linkage disequilibrium

The asymptotic approximation of the pseudo likelihood-ratio distribution breaks down when values of Z_a, Z_d are correlated due to LD. One way to overcome this is to 'prune' SNPs by hiererarchical clustering until only those with negligible correlation remain. A disadvantage with this approach is that it is difficult to control which SNPs are retained in an unbiased way without risking removal of SNPs which contribute greatly to the difference between subgroups.

I opted to use the LDAK algorithm [Speed et al., 2012], which assigns weights to SNPs approximately corresponding to their 'unique' contribution. Denoting by ρ_{ij} the correlation between SNPs i, j, and $d(i,j)$ their chromosomal distance, the weights w_i are computed so that

$$w_i + \sum_{i \neq j} w_j \rho_{ij}^2 e^{-\lambda d(i,j)} \tag{5.12}$$

is close to constant for all i, and $w_i > 0$ for all i. The motivation for this approach is that $\sum_{i \neq j} \rho_{ij}^2$ represents the replication of the signal of SNP i from all other SNPs.

This approach has the advantage that if n SNPs are in perfect LD, and not in LD with any other SNPs, each will be weighted $1/n$, reducing the overall contribution to the likelihood to that of one SNP. In practice, the linear programming approach results in many SNP weights

being 0. Using the LDAK algorithm therefore allows more SNPs to be retained and contribute to the model than would be retained in a pruning approach.

A second advantage of LDAK is that it homogenises the contribution of each genome region to the overall pseudo-likelihood. Many modern microarrays fine-map areas of the genome known or suspected to be associated with traits of interest [Cortes and Brown, 2011] which could theoretically lead to peaks in the distribution of SNP effect sizes, disrupting the assumption of normality. LD pruning and LDAK both reduce this effect by homogenising the number of tags in each genomic region.

I adapted the pseudo-likelihood function to the weights by multiplying the contribution of each SNP to the log-likelihood by its weight (equation 5.4.2), essentially counting the ith SNP w_i times over. Adjusting using LDAK was effective in enabling the distributions of PLR to be well-approximated by mixture-χ^2 distributions of the form 5.2 (appendix D.2, figures D.5a, D.5b and D.5c).

5.4.4 E-M algorithm to estimate model parameters

I use an expectation-maximisation algorithm [Dempster et al., 1977, Hastie et al., 2001] to fit maximum-PL parameters. Given an initial estimate of parameters $\Theta_0 = (\pi_1^0, \pi_2^0, \tau^0, \sigma_2^0, \sigma_2^0, \rho^0)$ I iterate three main steps:

1. Define for SNP s with Z scores $Z_d^{(s)}$,$Z_a^{(s)}$

$$\zeta_g^{(s)} = Pr(s \in \text{ category } g|\Theta_i)$$
$$\propto \begin{cases} \pi_1^i N_{\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)}(Z_d^{(s)}, Z_a^{(s)}) & (g=1) \\ \pi_2^i N_{\left(\begin{smallmatrix}1 & 0\\ 0 & (\sigma_2^i)^2\end{smallmatrix}\right)}(Z_d^{(s)}, Z_a^{(s)}) & (g=2) \\ \pi_3^i \left(\frac{1}{2} N_{\left(\begin{smallmatrix}(\tau^i)^2 & \rho^i\\ \rho^i & (\sigma_3^i)^2\end{smallmatrix}\right)}(Z_d^{(s)}, Z_a^{(s)}) + \frac{1}{2} N_{\left(\begin{smallmatrix}(\tau^i)^2 & -\rho^i\\ -\rho^i & (\sigma_3^i)^2\end{smallmatrix}\right)}(Z_d^{(s)}, Z_a^{(s)}) \right) & (g=3) \end{cases} \tag{5.13}$$

2. For $g \in (1,2,3)$ and LDAK weight w_s for SNP s set

$$\pi_g^{i+1} = \frac{\sum w_s \zeta_g^{(s)} + C}{\sum w_s + 3C} \tag{5.14}$$

3. Set

$$(\tau^{i+1}, \sigma_2^{i+1}, \sigma_3^{i+1}, \rho^{i+1}) = \arg\max_{(\tau,\sigma_2,\sigma_3,\rho)} PL(Z_d, Z_a | \pi_1^{i+1}, \pi_2^{i+1}, \tau, \sigma_2, \sigma_3, \rho) \tag{5.15}$$

Step 3 is complicated by the lack of closed form expression for the maximum likelihood estimator of ρ (because of the symmetric two-Gaussian distribution of category 3), requiring a bisection method for computation. The algorithm is continued until $|PLR(Z_d, Z_a|\Theta_i) - PLR(Z_d, Z_a|\Theta_{i-1})| < \varepsilon$; I used $\varepsilon = 1 \times 10^{-5}$.

The algorithm can converge to local rather than global minima of the likelihood. I overcome this by initially computing the pseudo-likelihood of the data at 1000 points throughout the parameter space, retaining the top 100, and dividing these into 5 maximally-separated clusters. The full algorithm is then run on the best (highest-PL) point in each cluster

An appropriate choice of Θ_0 can speed up the algorithm considerably; for simulations, I begin the algorithm at previous maximum-PL estimates of parameters for earlier simulations.

Maximum-cPL estimations of parameters were made using generic numerical optimisation with the *optim* function in R. Prior to applying the algorithm, parameters π_2 and σ_2 are estimated as maximum-PL estimators of the objective function

$$g(Z_a|\pi_2, \sigma_2) = \sum w_i \log\{(1-\pi_2)N_{0,1}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2} Z_a^{(i)}\} \tag{5.16}$$

where w_i is the weight for SNP i (see appendix C, section C.4 for rationale). The conditional pseudo-likelihood (*cPL*) was maximised over the remaining parameters.

The algorithm and other processing functions are implemented in an R package available at https://github.com/jamesliley/subtest

5.4.5 Properties and assumptions of the PLR test

My assumption that (Z_a, Z_d) follows a mixture Gaussian is generally reasonable for complex phenotypes with a large number of associated variants [Lo et al., 2015] and the adjustment for the distribution of Z_a (essentially conditioning on observed Z_a) reduces reliance on this assumption. If subgroup prevalence is unequal between the study group and population, my method can still be used with adaptation (appendix C, section C.2.4).

The test is robust to confounders arising from differential sampling to the same extent as conventional GWAS. For example, if subgroups were defined based on population structure,

and population structure also varied between the case and control group, SNPs which differed by ancestry would also appear associated with the disease, leading to a loss of control of type-1 error rate. However, the same study design would also lead to identification of spurious association of ancestry-associated SNPs with the phenotype in a conventional GWAS analysis. As for GWAS, this effect can be alleviated by including the confounding trait as a covariate when computing p-values (appendix C, section C.2).

5.4.6 Prioritisation of single SNPs

An important secondary problem to testing H_0 is the determination of which SNPs are likely to be associated with disease heterogeneity. I sought a way to test the association of a SNP with subgroup status (ie, Z_d), which gives greater priority to SNPs potentially associated with case/control status (ie, high Z_a).

An effective test statistic meeting these requirements is the Bayesian conditional false discovery rate (cFDR) [Andreassen et al., 2013] (see chapters 2, 3). In this context, the cFDR is used to test against the null hypothesis H_0' that the population minor allele frequencies of the SNP in both case subgroups are equal (ie, that the SNP does not differentiate subgroups), responding to association with case/control status in a natural way by relaxing the effective significance threshold on $|Z_d|$. As discussed in chapter 2, this relaxation of threshold only occurs if there is systematic evidence that high $|Z_d|$ scores and high $|Z_a|$ scores typically co-occur. The cFDR is direction-independent, only considering p-values.

Given a set of observed Z_a and Z_d values $Z_a^{(i)}$, $Z_d^{(i)}$, with corresponding two-sided p values p_{ai}, p_{di}, the cFDR for SNP j is defined as

$$\begin{aligned} X_4 &= p_{dj}\frac{\left|\{i : p_{ai} \leq p_{aj} \wedge p_{di} \leq p_{dj}\}\right|}{\left|\{i : p_{di} \leq p_{dj}\}\right|} \\ &\approx Pr(H_0' \big| P_a \leq p_{aj}, P_d \leq p_{dj}) \end{aligned} \tag{5.17}$$

A false-discovery rate bound can be estimated on such SNPs as per chapter 2.

I discuss three other single-SNP test statistics in appendix C, section C.6.1, which test against different null hypotheses. If the hypothesis H_0' is to be tested, then I consider the cFDR the best of these.

Contour plots of the test statistics for several datasets are shown in appendix D.2, figures D.8, D.9 and D.10.

5.4.7 Genetic correlation testing

Given the correlation between Z_d and Z_a in the age-at-diagnosis analysis, methods to estimate narrow-sense genetic correlation (r_g) [Bulik-Sullivan et al., 2015, Lee et al., 2012] may be adaptable to the subgrouping question by estimating r_g across a set of SNPs between case/control traits of interest, with the potential advantage of characterising heterogeneity using a single widely-interpretable metric. This may be between Z scores derived from comparing the control group to each case subgroup, testing under the null hypothesis $r_g = 1$ (method 1); or between the familiar Z_a and Z_d, under the null hypothesis $r_g = 0$ (method 2).

I explored these methods in appendix C, section C.5. I show that method 1 leads to systematically high false positive rates, as r_g is also reduced from 1 in subgroupings that are independent of the overall disease process (e.g. hair colour in T2D). I show that method 2 is considerably less powerful than my method because it tests a narrower definition of H_1 which does not take account of the marginal variances of the distribution of Z_d, Z_a in category 3, and requires that correlation between Z_d and Z_a be always positive or always negative, in contrast to my symmetric model (Figure 5.1). Indeed, parameter ρ estimates an analogue of r_g accounting for simultaneous correlation and anticorrelation.

Methods to compute r_g were not explicitly proposed as methods for subgroup testing, and my analysis does not indicate any general shortcomings. However, comparison with r_g based approaches places my method in the context of established methodology, demonstrating the necessity of considering both variance parameters (τ, σ_3) and covariance parameters (ρ) in testing a subgrouping of interest.

5.4.8 Description of GWAS datasets

ATD samples were genotyped on the ImmunoChip [Cortes and Brown, 2011] a custom array targeting putative autoimmune-associated regions. Data were collected for GWAS-like analyses of dense SNP data [Cooper et al., 2012]. The dataset comprised 2282 cases of Graves' disease, 451 cases of Hashimoto's thyroiditis, and 9365 controls.

T1D samples were genotyped on either the Illumina 550K or Affymetrix 500K platforms, gathered for a GWAS on T1D [Barrett et al., 2009]. I imputed between platforms in the same way as the original GWAS. The dataset comprised genotypes from 5908 T1D cases and 8825 controls, of which all had measured values of TPO-Ab, 3197 had measured IA2-Ab, 3208 had measured GAD-Ab, and 2240 had measured PCA-Ab. Comparisons for each autoantibody were made between cases positive for that autoantibody, and cases not positive for it. I did not attempt to perform comparisons of individuals positive for different autoantibodies (for

instance, TPO-Ab positive vs IA2-Ab positive) because many individuals were positive for both.

To generate summary statistics corresponding to geographic subgroups, I considered the subgroup of cases from each of twelve regions and each pair of regions against all other cases (78 subgroupings in total). To maximise sample sizes, I considered T1D cases as 'controls' and split the control group into subgroups.

5.4.9 Quality control

Particular care had to be taken with quality control, as Z-scores had to be relatively reliable for all SNPs assessed, rather than just those putatively reaching genome-wide significance. For the T1D/T2D/RA comparison, which I re-used from the WTCCC, a critical part of the original quality control procedure was visual analysis of cluster plots for SNPs reaching significance, and systematic quality control measures based on differential call rates and deviance from Hardy-Weinberg equilibrium (HWE) were correspondingly loose [The Wellcome Trust Case Control Consortium, 2007]. Given that we were not searching for individual SNPs, this was clearly not appropriate for my method.

I retained the original call rate (CR) and MAF thresholds (MAF $\geq 1\%$, CR $\geq 95\%$ if MAF $\geq 5\%$, CR $\geq 99\%$ if MAF <5%) but employed a stricter control on Hardy-Weinberg equilibrium, requiring $p \geq 1 \times 10^{-5}$ for deviation from HWE in controls. I also required that deviance from HWE in cases satisfied $p \geq 1.91 \times 10^{-7}$, corresponding to $|z| \leq 5$. The looser threshold for HWE in cases was chosen because deviance from HWE can arise due to true SNP effects [Anderson et al., 2010]. I also required that call rate difference not be significant ($p \geq 1 \times 10^{-5}$) between any two groups, included case-case and case-control differences. Geographic data was collected by the WTCCC and consisted of assignment of samples to one of twelve geographic regions (Scotland, Northern, Northwestern, East and West Ridings, North Midlands, Midlands, Wales, Eastern, Southern, Southeastern, and London [The Wellcome Trust Case Control Consortium, 2007]). In analysing differences between autoimmune diseases, I stratified by geographic location; when assessing subgroups based on geographic location, i did not.

For the ATD and T1D data, I used identical quality control procedures to those employed in the original paper [Cooper et al., 2012, Barrett et al., 2009]. I applied genomic control [Devlin et al., 2001] to computation of Z_a and Z_d scores except for the analysis of ATD (following the original authors [Cooper et al., 2012]) and the geographic analyses (as

discussed above). In all analyses except where otherwise indicated I removed the MHC region with a wide margin ($\approx 5Mb$ either side).

5.5 Addendum: alternative methods for testing

5.5.1 Introduction

A major shortcoming of the PLR method is its complexity, in regard to both the number of steps involved in using the method, and the computational resources required. The determination of whether the joint distribution of (Z_d, Z_a) contains a component with both marginal variances > 1 appears relatively simple, to the point where it appears to be testable more simply. The main text in this chapter and appendix C discuss the advantages and shortcomings of using genetic correlation in several ways as a substitute for the PLR. In this addendum, I discuss several other alternatives to testing the main hypothesis, and propose a custom test statistic which may be more effective. This section is a description of a promising new direction for future investigation rather than a description of completed work.

A generalised version of the null hypothesis H_0^G to be tested against is as follows:

$$\begin{aligned} H_0^G : PDF_{Z_d,Z_a}(x,y) = \pi_1 N_{I_2}(x,y) \\ + \pi_2 N_1(x) F(y) \\ + \pi_3 N_1(y) G(x) \end{aligned} \tag{5.18}$$

where F and G are univariate density functions which may have variance > 1, and $\pi_1 + \pi_2 + \pi_3 = 1$. In general, F and G are considered to be continuous, bounded and strictly positive in $\mathbb{R}$. The three components correspond to the three categories used throughout the chapter. Notably, Z_d and Z_a are conditionally independent given the category. Denote by $H_0^G(\pi_1), H_0^G(\pi_1, F), H_0^G(\pi_1 = x), H_0^G(F = f(x))$... the subspaces of H_0^G with the appropriate parameters fixed, taking the specified values if given.

The possibility of attempting to reject H_0^G (or more specific null hypotheses) by considering correlation between Z_a and Z_d is discussed at length in section 5.4.7. The problems with using standard correlation or 'absolute correlation' are summarised in table D.1 in appendix D.1.

Linear functions of genotypes (genetic risk scores) can not generally reject H_0^G

Another potential approach to rejecting H_0^G is to use genetic risk scores (GRS). If a predictive score were developed to differentiate cases from controls, H_0^G could be tested by assessing the degree to which the score can differentiate the disease subgroups. A similar approach is used in chapter 6 in understanding the subtypes of JIA.

While effective in some circumstances, GRS are also unsuitable for the problem in many cases. Consider an generalised GRS to predict a genotype Y which for sample i takes the form

$$GRS(i) = \sum_{j \in SNPs} \beta_j g_{ij} \tag{5.19}$$

where $g_{ij} \in G_j \in G$ is the genotype for sample i, SNP j, and $\beta_j = f_j(G,Y)$ is a function dependent on all genotypes and all case/control labels, but independent of subtype labels.

Denote the case group by C_{12}. Suppose that the disease consists of two subtypes S_1 and S_2 of equal prevalence, where subgroups have completely distinct genetic associations (this scenario is in H_1), but for which the distribution of effect sizes is the same. For a SNP s_1 associated with S_1 and a SNP s_2 associated with S_2 of the same MAF and effect size, the coefficients β_j in the GRS (equation 5.19) will have the same expected value, and indeed the same distribution.

Since the subgroups are of equal prevalence and have the same effect size distribution, a set of observed genotypes G' with subtype labels Z' has the same likelihood as the same set G' with subtype labels swapped. Thus, if the GRS in equation 5.19 is used to differentiate subgroups of C_{12} corresponding to S_1 and S_2, and the difference in mean GRS score is

$$\Delta G(S_1,S_2) = \frac{1}{|\{i \in S_2\}|} \sum GRS(i|i \in S_2) - \frac{1}{|\{i \in S_1\}|} \sum GRS(i|i \in S_1) \tag{5.20}$$

then

$$\begin{aligned} E\left(\Delta G(S_1,S_2)\right) &= E\left(\Delta G(S_2,S_1)\right) \\ &= -E\left(\Delta G(S_1,S_2)\right) \\ &= 0 \end{aligned} \tag{5.21}$$

so the GRS will not be able to differentiate subgroups in this case. A criterion based on whether GRS fitted to case/control status can differentiate subgroups forms a sufficient but not necessary condition for falsehood of H_0^G.

5.5.2 Custom statistic

The main difficulty in finding a test statistic for rejecting H_0^G above is determining a function that has an estimable distribution and expected value under H_0^G but whose behaviour changes under the H_1 scenarios in table D.1. Effectively, SNPs should contribute more to the test statistic if *both* $|Z_d|$ and $|Z_a|$ are large. A SNP with low $|Z_d|$ is of relatively little help in rejecting H_0, whatever its value of Z_a.

An ideal metric $k(Z_d, Z_a)$ to identify such pairs of (Z_d, Z_a) scores should thus increase as both Z_d and Z_a increase, but have a limiting maximum value as $Z_a \to \infty$ with Z_d fixed. The metric $k(Z_d, Z_a) = \min(|Z_d|, |Z_a|)$ is one choice, but it is too extreme; there should be *some* variance in the 'interest' in a (Z_d, Z_a) score with Z_a when $|Z_a| \gg |Z_d|$. On the X-Y plane, $k(x, y)$ should have contours roughly resembling plots of the function:

$$(x^2 - k)(y^2 - k) = a^2 \tag{5.22}$$

with a held constant, and k varying (shown in figure 5.6.

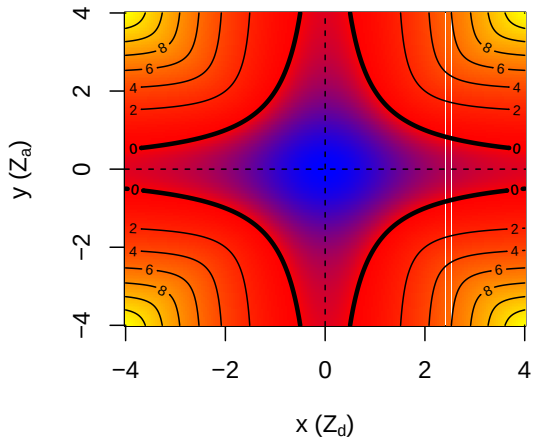


Fig. 5.6 Plot of implicitly defined curves $(x^2 - k)(y^2 - k) = a^2$ for different values of k (shown as contour labels and colours, with blue and yellow indicating negative and positive values of k respectively). While potentially useful as a metric for 'interest' in a (Z_d, Z_a) pair in the subgroup context, a problem is the variable behaviour when $Z_d = 0$ or $Z_a = 0$.

This immediately suggests the metric

$$k(x,y) = x^2 + y^2 - \sqrt{(x^2-y^2)^2 + (2a)^2} \tag{5.23}$$

A problem with this metric is that it is negative at and around the origin, but tends to 0 as $Z_d \to \infty, Z_a = 0$ and as $Z_a \to \infty$, $Z_d = 0$. This is a problem, as (Z_d, Z_a) pairs with one value zero are equally (un)interesting everywhere in this context. I thus propose the non-negative function $k(x,y)$ parametrised by a (in the place of $(2a)$ above) as

$$k(x,y) = \sqrt{x^4 + a^2} + \sqrt{y^4 + a^2} - \sqrt{(x^2-y^2)^2 + a^2} - a \tag{5.24}$$

Defining $Z_d^{(j)}$, $Z_a^{(j)}$ as the Z_d and Z_a scores at a SNP j, a potential test statistic K for the subgrouping problem is then defined as a function of the sets of values $\{Z_d^{(j)}\}$ and $\{Z_a^{(j)}\}$

$$K(\{Z_d^{(j)}\}, \{Z_a^{(j)}\}) = \frac{\sum w_j k(Z_d^{(j)}, Z_a^{(j)})^\gamma}{\sum w_j} \tag{5.25}$$

where the sum is over all SNPs j. Weights w_j may be derived from LDAK or similar procedures. The parameter a effectively controls how curved the contours of k are near the origin (see figure 5.7) and the parameter γ controls the degree to which individual variants contribute to the test statistic, and the parameter a . A contour plot of $k(Z_d, Z_a)$ is shown in figure 5.7.

The function k has many of the desired properties of a metric for 'interest' in (Z_d, Z_a) pairs. Firstly, since with x fixed,

$$\lim_{y\to\infty} \left(\sqrt{y^4 + a^2} - \sqrt{(x^2-y^2)^2 + a^2} \right) = x^2 \tag{5.26}$$

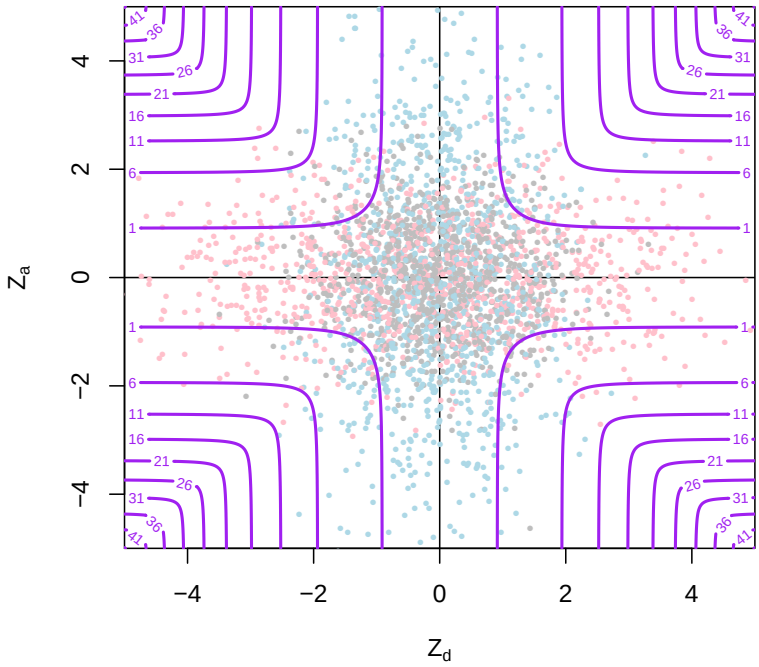


Fig. 5.7 Contour plot of alternative metric k for subgrouping problem, with an example distribution of Z_d, Z_a in H_0. Contours of $k(Z_d, Z_a)$ (with $a = 2$) are shown in purple, with an example of a Z_d, Z_a in H_0^G shown in grey, red and blue corresponding to categories 1, 2 and 3. The new test statistic differentiates points for which *both* Z_d and Z_a differ substantially from 0.

If H_0^G is restricted to well-behaved instances where $Z_d|Z_a = y, H_0^G$ converges surely to $N(0,1)$ as $y \to \infty$ (equation 5.18), we have

$$\begin{aligned}\lim_{y\to\infty} E\left\{k(Z_d,Z_a)|Z_a = y, H_0^G\right\} &= E\left\{\sqrt{Z_d^4 + a^2}|H_0^G\right\} \\ &\quad + \lim_{y\to\infty} E\left\{\sqrt{y^4+a^2} - \sqrt{\left(Z_d^2 - y^2\right)^2 + a^2} - a|H_0^G\right\} \\ &= E\left\{\sqrt{Z_d^4 + a^2}|H_0^G\right\} + E\left\{Z_d^2|H_0^G\right\} - a \\ &< \infty \end{aligned} \tag{5.27}$$

and indeed, the random variable $k(Z_d, Z_a)|Z_a = y, H_0^G$ converges surely as $y \to \infty$ (with appropriate conditions on F, G in equation 5.18) to a random variable ψ_a:

$$\psi_a = \sqrt{X^4 + a^2} + X^2 - a;\ X \sim N(0,1) \tag{5.28}$$

For $y > 0$

$$\begin{aligned} \frac{\delta}{\delta y} k(x,y) &= \frac{2y^3}{\sqrt{y^4 + a^2}} + \frac{2y(x^2 - y^2)}{\sqrt{(x^2 - y^2)^2 + a^2}} \\ &> 0 \end{aligned} \tag{5.29}$$

so the value of $k(x,y)$ increases as y increases with fixed x. Hence for $y_1 > y_2 > 0$ we have $\mathbf{1}_{k(x,y_1)>k_0} < \mathbf{1}_{k(x,y_2)>k_0}$. Denote $H_0' = H_0^G(\pi_1, \pi_2, F, G)$ and $f_{Z_d}(x)$ the PDF of Z_d under H_0'. Now

$$\begin{aligned} Pr(k(Z_d, Z_a) > k_0 | Z_a = z_0, H_0') &= \int_{\mathbb{R}} \mathbf{1}_{k(x,y_1)>k_0} f_{Z_d}(x) dx \\ &< \lim_{y \to \infty} \int_{\mathbb{R}} \mathbf{1}_{k(x,y)>k_0} f_{Z_d}(x) dx \\ &= \lim_{y \to \infty} Pr(k(Z_d, Z_a) > k_0 | Z_a = y, H_0') \\ &= Pr(\psi_a > k_0) \end{aligned} \tag{5.30}$$

Since this is independent of f_{Z_d}, the same holds replacing H_0' with H_0^G. The same also holds with Z_d and Z_a interchanged.

Define random variables ψ_a as above and Ψ_a:

$$\Psi_a = k(X,Y);\ \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \tag{5.31}$$

If $\psi_a = \sqrt{X^4 + a^2} + X^2 - a = k_0$ then $X = \pm\sqrt{\frac{k(2a+k)}{2(a+k)}}$. The region $\{(x,y) : k(x,y) \leq k_0\}$ is thus contained within the region $\{(x,y) : |x| \geq \sqrt{\frac{k(2a+k)}{2(a+k)}}, |y| \geq \sqrt{\frac{k(2a+k)}{2(a+k)}}\}$ and for all k_0

$$\begin{aligned} Pr(\Psi_a > k_0) &< (Pr(\psi_a > k_0))^2 \\ &= 4\Phi\left(-\sqrt{\frac{k(2a+k)}{2(a+k)}}\right)^2 \\ &< 2\Phi\left(-\sqrt{\frac{k(2a+k)}{2(a+k)}}\right) \end{aligned} \tag{5.32}$$

PDFs and CDFs for ψ_a and Ψ_a are shown in figure 5.8. For a specific instance $H_0' =$

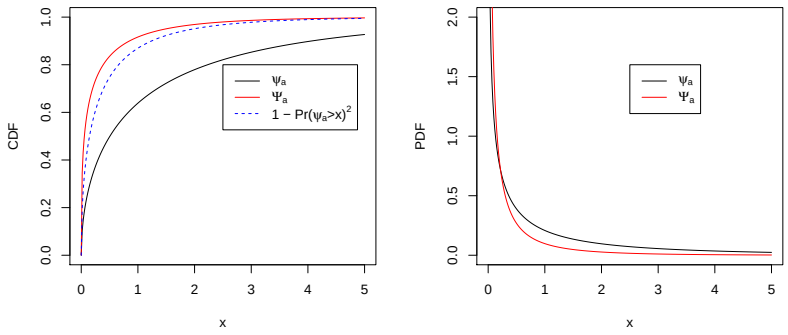


Fig. 5.8 CDFs and PDFs for random variables ψ_a and Ψ_a. CDF plot additionally shows $1 - Pr(\psi_a > x)^2$ which is strictly less than $Pr(\Psi_a < x)$.

$H_0(\pi_1,\pi_2,F,G)$ of H_0 with π_1,π_2,F,G fixed,

$$\begin{aligned}
Pr(k>k_0|H_0') &= \iint_{\mathbb{R}^2} \mathbf{1}_{k(x,y)>k_0} PDF(Z_d,Z_a)(x,y)dxdy \\
&= \pi_1 \iint_{\mathbb{R}^2} \mathbf{1}_{k(x,y)>k_0} N_{I_2}(x,y)dxdy \\
&\quad + \pi_2 \iint_{\mathbb{R}^2} \mathbf{1}_{k(x,y)>k_0} N_1(x)F(y)dxdy \\
&\quad + \pi_3 \iint_{\mathbb{R}^2} \mathbf{1}_{k(x,y)>k_0} N_1(y)G(x)dxdy \\
&< \pi_1 Pr(\Psi_a>k_0) + (\pi_2+\pi_3)Pr(\psi_a>k_0) \qquad (5.33) \\
&< 4\pi_1\Phi\left(-\sqrt{\frac{k(2a+k)}{2(a+k)}}\right)^2 + 2(1-\pi_1)\Phi\left(-\sqrt{\frac{k(2a+k)}{2(a+k)}}\right) \\
&< 2\Phi\left(-\sqrt{\frac{k(2a+k)}{2(a+k)}}\right) \qquad (5.34)
\end{aligned}$$

The bound 5.34 does not depend on π_1,π_2,F,G and hence the bound holds when H_0' is replaced by H_0. The bound 5.33 only depends on π_1, and hence H_0' may be replaced by $H_0^G(\pi_1)$. The distributions of ψ_a and Ψ_a are empirically tractable, enabling construction of a rough bound on a p-value associated with K under H_0^G or $H_0^G(\pi_1)$ using the central limit theorem.

Distribution of new test statistic under H_0

The bound introduced in equation 5.34 is not very useful in practice, as it is usually sufficient to reject weaker forms of H_0^G. It is generally reasonable to make some assumptions on the values π_1,π_2,F,G can take; this is effectively the approach taken with the PLR, in which the null hypothesis has specified parameters $(\pi_1,\pi_2,\pi_3,\tau,\sigma_1,\sigma_2=1,\rho=0)$. Given an estimator $\hat{\pi_1}$ of π_1, then bound 5.33 with $\hat{\pi_1}$ in place of π_1 can be used to accept or the null hypothesis $H_0^G(\pi_1=\hat{\pi_1})$. Since for any $\pi_1'<\hat{\pi_1}$ we have

$$Pr\left(k>k_0|H_0^G(\pi_1=\pi_1')\right) < Pr\left(k>k_0|H_0^G(\pi_1=\hat{\pi_1})\right) \qquad (5.35)$$

a test against the null hypothesis $H_0^G(\pi_1=\hat{\pi_1})$ is conservative against $H_0^G(\pi_1=\pi_1')$ in that the true p-value is necessarily lower. Thus for a conservative test, the value for $\hat{\pi_1}$ should be an upper bound on the true value.

Further improvements to the test can be made if estimates are made of the distributions F and G. Importantly, the estimation of these distributions only requires a one-dimensional model; under H_0, the marginals PDFs of Z_d and Z_a are

$$\begin{aligned} PDF_{Z_d}(x) &= (\pi_1 + \pi_2)N_1(x) + \pi_3 G(x) \\ PDF_{Z_a}(x) &= (\pi_1 + \pi_3)N_1(x) + \pi_2 F(x) \end{aligned} \tag{5.36}$$

suggesting the following procedure:

1. Estimates $\hat{\pi_3}$ and $\hat{G}$ of π_3 and (parameters of) G are made from observed values of Z_d
2. Estimates $\hat{\pi_2}$ and $\hat{F}$ of π_2 and (parameters of) F are made from observed values of Z_a
3. The CDF of K under the $H_0^s = H_0(\pi_1 = \hat{\pi_1}, \pi_2 = \hat{\pi_2}, \pi_3 = \hat{\pi_3}, F = \hat{F}, G = \hat{G})$ is computed, generally empirically
4. An observed value of K corresponding to a subtyping of interest is tested for consistency with H_0^s by comparison to this empirical CDF

This procedure is a promising avenue for future work on this test statistic, but was not investigated in this instance.

5.5.3 Comparison of power of new method and PLR

Application to autoimmune datasets

I tested whether the new method could reject H_0^G or $H_0^G(\pi_1 = \hat{\pi_1})$ on the autoimmune datasets used to test the PLR in section 5.2.3 (considering each pair of phenotypes in (T1D,T2D,RA) as subtypes of a wider autoimmune phenotype). These results should be interpreted cautiously in the absence of simulations to check for appropriate type-1 error rate control.

Firstly, I computed the value of K for each set of Z_d and Z_a scores as per equation 5.25, using LDAK weights, considering values of γ in $(1,2,4)$. I then considered two different models for the distribution of $k(Z_d, Z_a)$ under H_0, as follows:

1. $k(Z_d, Z_a)|H_0^G \sim \psi_a$ (bound 5.34)
2. $k(Z_d, Z_a)|H_0^G(\pi_1 = \hat{\pi_1}) \sim \begin{cases} \psi_a & \text{prob. } 1-\pi_1 \\ \Psi_a & \text{prob. } \pi_1 \end{cases}$ (bound 5.33)

In model 2, the value $\hat{\pi}_1$ was estimated from the observed distribution of Z_d, Z_a using steps 1 and 2 and setting $\hat{\pi}_1 = 1 - \hat{\pi}_2 - \hat{\pi}_3$.

The null hypothesis H_0^G could not be rejected in any case using model 1 ($p > 0.5$ in all cases). Confidence in rejecting $H_0^G(\pi_1)$ using model 2 varied according to γ, with p-values shown in table 5.3. Performance was markedly strongest at $\gamma = 4$.

Table 5.3 P-values for rejection of $H_0^G(\pi_1)$ using new test statistic on datasets for T1D/RA, T1D/T2D, T2D/RA, and GD/HT. See table 5.2 for comparison.

Subgroups	$\gamma = 1$	$\gamma = 2$	$\gamma = 4$
T1D/RA	> 0.5	0.34	2×10^{-3}
T1D/T2D	> 0.5	7×10^{-4}	$< 1 \times 10^{-4}$
T2D/RA	> 0.5	2×10^{-3}	$< 1 \times 10^{-4}$

Simulations

In order to establish a comparison of the potential power of the new statistic with the power of the PLR method, I re-simulated and re-tested the scenarios discussed in the main paper. In each case, these reflect the power of the new statistic to reject $H_0^G(\pi_1, \pi_2, F, G)$ rather than H_0^G, where F and G are normal distributions with mean 0. In this sense, the power in these simulations represents the 'optimal' power of the new test statistic.

For each set of parameters $(\pi_1, \pi_2, \pi_3, \tau, \sigma_1, \sigma_2, \rho)$ I simulated 5000 sets of Z_d and Z_a scores under the full model and the null model (with σ_2 replaced by 1 and ρ by 0). I then estimated the power of the new test to reject the null hypothesis at $\alpha = 0.05$ by computing the proportion of simulations from the full model for which the associated test statistic K passed the 95% quantile of test statistics K from simulations from the null model. For these simulations, I used $\gamma = 1$; brief analyses suggested similar behaviour when γ was set to 2 or 4.

A plot of power in the same circumstances as for the PLR is shown in figure 5.9, and more extensive plots (again in the same circumstances as for PLR) are shown in appendix D.2, figures D.4.

Figure 5.10 compares power between the PLR and new method at all points of the space H_1 considered in appendix D.2, figures D.4. In general, the PLR tends to be more powerful, although not in all circumstances. The greatest difference in power in favour of K over PLR is in cases where π_3 is high and τ is low, corresponding to a subgrouping trait which is highly polygenic but with small effect sizes. In general, the power of K is sufficient that it

appears usable as a test statistic, with the advantage of being far simpler than the PLR though typically with somewhat lower power.

5.5.4 Discussion

The major areas for further work on this statistic are determination of an appropriate null hypothesis and development of methods to estimate the distribution of K under this null, whether by simulation or analytically. The null hypothesis proposed in equation 5.18 is too broad to realistically use in practice. Other areas for consideration are the choice of value for the parameters a and γ, and the management of situations where Z_d, Z_a are systematically correlated (for example, see appendix C, section C.2.4).

The major reason why the PLR-testing procedure is so complex is the difficulty in effectively simulating under H_0 by permutation testing or otherwise. Namely, it is relatively easy to simulate with $\tau = 1$, but very difficult to simulate with $\tau > 1$. It may be possible to avoid the need for permutation testing using the new method, but this needs to be tested more extensively.

The new test statistic is often less powerful than the PLR, but not universally, and it may be well-suited to assessing classes of phenotypes for which the genetic architecture of the subgrouping trait involves many small effects. One example may be analysis of subgroups corresponding to geography [Leslie et al., 2015]. Overall, the new statistic appears to be a reasonable alternative option to testing in the subgrouping problem, which has the advantage of being considerably simpler than the PLR. It may serve as a useful indicator for certain types of disease heterogeneity, and contribute to the understanding of inter-phenotypic genetic relationships.

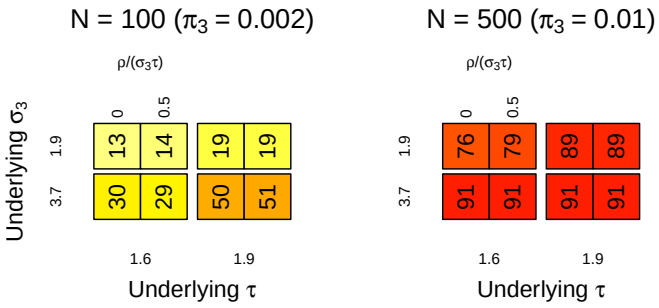


Fig. 5.9 Power of new test statistic K depends on the number of SNPs in category 3 and the underlying values of model parameters σ_2, σ_3, τ, ρ. Dependence on number of case/control samples arises through the magnitudes of σ_3 and τ (appendix C, section C.4.3). Figure shows power estimates for various values of π_3, σ_3, τ, ρ. Value N is the approximate number of SNPs in category 3, ($\propto \pi_3$). Each simulation was on 5×10^4 simulated autosomal SNPs in linkage equilibrium. Value $\rho/(\sigma_3\tau)$ is the absolute correlation between Z_d and Z_a in category 3. Also see appendix D.2, figure D.4.

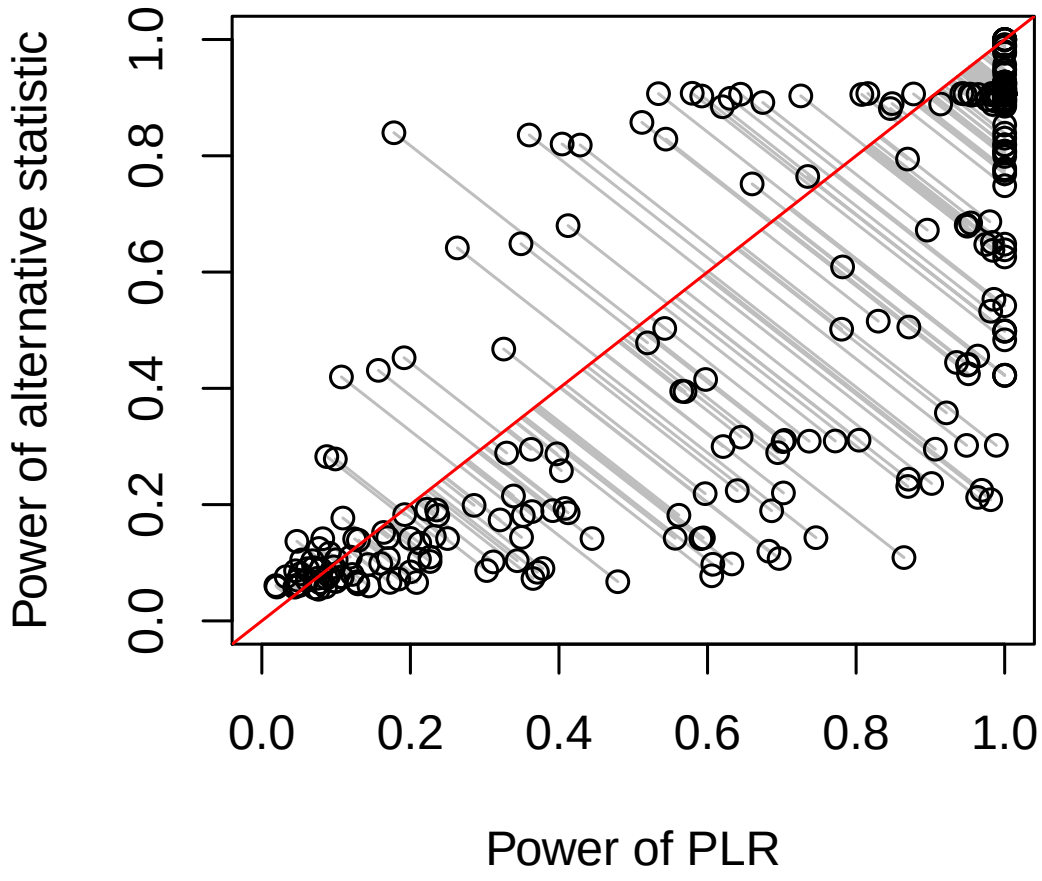


Fig. 5.10 Comparison of power of PLR and power of new test statistic K to reject $H_0^G(\pi_1, \pi_2, F, G)$ at $\alpha = 0.05$. Lines are drawn from each point to the x-y line to indicate deviation from equal power. In general, the PLR method is more powerful, but this is not universal.

Chapter 6

Levering phenotypes for risk prediction

6.1 Introduction

Following chapter 3, Juvenile Idiopathic Arthritis (JIA) is a clinically-diagnosed disease characterised by a persistent arthritis in children in the absence of other known causes. As a 'diagnosis of exclusion', JIA encompasses all childhood arthritis of unknown cause and is therefore probably caused by a range of pathological processes. The standard clinical sub-classification of JIA was proposed by the International League of Associations for Rheumatology (ILAR) [Petty et al., 2004] with the understanding that subtype definitions and sub-classifications would evolve as further information on the disease became available [Ravelli and Martini, 2007].

The ILAR subtypes are based on the number of joints affected (oligoarthritis: ≤ 4; polyarthritis: > 4), seropositivity to rheumatoid factor (RF), and the presence of extra-articular symptomatology (systemic (Still's disease), enthesitis-related, and psoriatic arthritis) [Ravelli and Martini, 2007], with a proportion of cases remaining undifferentiated. Population frequencies of each subtype within JIA as a whole are shown in table 6.1. The oligoarthritis

Table 6.1 Prevalence of subtypes of JIA. Reproduced from [Ravelli and Martini, 2007]

Subtype	**Abbreviation**	**Frequency (%)**
Systemic arthritis	Sys	4–17
Oligoarthritis	EO/PO	27–56
RF-positive polyarthritis	RF+poly	2–7
RF-negative polyarthritis	RF-poly	11–28
Enthesitis-related arthritis	ERA	3–11
Psoriatic arthritis	PSA	2–11

subtype can be further sub-categorised into 'extended' (EO) and 'persistent' (PO) subtypes based on whether the patient remains in an oligoarthritic state or the disease progresses to polyarthritis [Gardner-Medwin et al., 2006], an assessment which can obviously not be made at the time of initial diagnosis of the disease.

Due the heterogeneity and relative commonality of JIA, the overall genetic architecture would be expected to be polygenic; this is supported by GWAS/ImmunoChip findings [Hinks et al., 2013] showing a widespread distribution of associations across the genome (also see chapter 3). An important consideration in the genetic analysis of JIA is the individual genetic architectures of each subtype. There may be genetic heterogeneity amongst patients with different JIA subtypes in that the individual genetic architectures may be distinct to the genetic architecture of JIA as a whole; indeed, this may be expected given the clinical heterogeneity. More generally, subtype status may be driven by different pathological processes, be driven by individual anatomical or physiological variation, or be driven by environmental differences (see chapter 5 and table D.1 in appendix D.1), with the first two phenomena being of particular scientific interest.

The systemic subtype of JIA (Sys) appears to have a genetic architecture largely distinct from that of other subtypes. In an international GWAS of 770 children with Sys JIA [Ombrello et al., 2017], genetic risk scores (GRS) trained to non-systemic JIA could not predict Sys cases from controls, and a GRS trained to the RF-positive polyarthritis found Sys samples to be significantly differentiated from control samples in the opposite direction to RF-positive polyarthritis samples. The differential genetic architectures highlight the difficulties in analysing diseases with subtypes of this sort, in that the analysis of systemic JIA is unlikely to be strengthened by levering on general JIA vs control datasets. It is likely that further genetic heterogeneity is present between JIA subtypes, although this is difficult to determine using genome-wide significant SNPs only [Hinks et al., 2013]. Understanding the genetic heterogeneity of JIA subtypes may be useful both in the development of specific treatment and in direct clinical management, and could be used to inform more clinically-useful sub-classifications of the disease in the future [Petty et al., 2004, Ravelli and Martini, 2007].

In this chapter I seek to develop and apply methods for assessing genetic heterogeneity in JIA. The number of JIA samples in total is small (< 3000) and each subtype is sufficiently rare that it is difficult to study in isolation. Given the limited power of analyses on these sample sizes, my analyses typically involve leverage on summary statistics from higher-powered studies. An obvious candidate phenotype to lever on is the JIA/control phenotype, in the manner of the single-SNP analysis in chapter 5, section 5.2.2. The results from systemic JIA above suggest that this will not be useful for understanding the basis of Sys, and this

may also be the case for other subtypes. For this reason, I also lever on summary statistics other genetically related autoimmune diseases.

The most obvious adult analogue of JIA is RA, with the polyarthritic subtypes of JIA in particular sharing pathological characteristics of RA including symmetric distribution of infected joints and elevation of rheumatoid factor [Gardner-Medwin et al., 2006, Ravelli and Martini, 2007]. This suggests examination of the genetics of RA as a useful starting point for determining differential genetics of JIA subtypes. A second candidate for a discriminating GRS is T1D, which shows considerable pleiotropy with JIA as explored in chapter 3; indeed, as discussed in section 3.2.1, among a range of autoimmune diseases, JIA is the most similar disease to T1D on the basis of enrichment of summary statistics [Burren et al., 2014][1]. I sought to establish whether GRSs to predict disease/control status for RA and T1D could differentiate JIA subtypes from controls to different degrees.

An example of particular importance is the prediction of extended/persistent oligoarthritis state; children for whom extension of an existing oligoarthritis is unlikely could be treated less aggressively than those for whom extension is probable. Resources spent in follow-up of children with oligoarthritis could similarly be directed in a more efficient way.

In light of these clinical and scientific applications, this chapter has two main parts[2] :

1. Assessment of whether EO and PO subtypes of JIA can be differentiated on the basis of genetics, using GRS based on genotype data for EO/PO samples and other JIA samples, and summary statistics for RA and T1D.

2. Assessment of the similarity of genetic architecture of JIA subtypes to the genetic architecture of RA and T1D, using GRS based on genetic data from JIA and summary statistics for RA and T1D.

[1]The similarity between JIA and T1D in the absence of obvious phenotypic overlap is indicative of some uncertainty in predicting genetic overlap by phenotypic overlap, somewhat undermining the idea of levering on phenotypically related phenotypes

[2]The sections are ordered to reflect the chronological order of this work. The primary aim of this exercise was that in item 1 (discrimination of EO/PO), with item 2 (wider analysis of subtypes) as a later exploratory analysis. I am aware that if the order of this work had been reversed; that is, if the specific analysis of EO/PO differentiation had been incentivised by the results of the wider subtype analysis, then it would be necessary to perform appropriate multiple-testing adjustments on the EO/PO analysis. However, this was not the case, and multiple-testing adjustments are not used in item 1

6.2 Common considerations

6.2.1 Notation

Throughout this chapter, subscript i will generally refer to samples and be assumed to run from 1 to n unless otherwise specified. Subscript j will refer to variables/predictors and run from 1 to p. Subscripts k,h will refer to cross-validation folds and run from 1 to the number of folds. A set of values $\{X\}$ will be assumed to take the form of a column vector when appropriate.

6.2.2 Datasets

The JIA dataset consisted of 2585 JIA cases and 5181 controls from a range of recruitment centres across the UK. The dataset was originally gathered for a GWAS on JIA, currently in preparation. I was not involved in data gathering, sample recruitment or preparation, or quality control procedures. The breakdown of JIA samples is shown in table 6.2. Specifically,

Table 6.2 Number of available cases in each JIA subtype

Subtype	**N cases**	**Frequency (%)**
Sys	283	10.9
EO	394	15.2
PO	650	25.1
RF+poly	199	7.7
RF-poly	573	22.2
ERA	185	7.2
PsA	150	5.8
Unclassified	86	3.3
Missing	65	2.5
Control	5833	-

the dataset included 1044 cases initially presenting with oligoarthritis, of which 650 (62%) progressed to polyarthritis. Samples were genotyped on a range of platforms and imputed to approximate genome-wide cover. SNPs were removed if they were multi-allelic, had imputation $r^2 < 0.5$, had minor allele frequency $< 1\%$, or deviated from Hardy-Weinberg equilibrium in controls with p-value $< 1 \times 10^{-3}$. Sample-wise quality control measures were identical to those described in [Ombrello et al., 2017]. After quality control, 7292621 SNPs (including imputed variants) were available for analysis.

Datasets used for leverage included summary statistics from GWAS on rheumatoid arthritis [Okada et al., 2014] and type 1 diabetes [Barrett et al., 2009]. Summary statistics in

all cases were the same as those reported in the final publications, following quality control procedures as described therein. Since 'cases' and 'controls' both referred to JIA samples in this case, there were no shared samples between the datasets used for leverage and the JIA dataset used for GRS training.

6.2.3 Genetic risk scores

I briefly introduced the concept of genetic risk scores in chapter 1, section 1.3.5. In its simplest form, a genetic risk score *GRS* for a sample i with numeric genotypes g_{ij} at SNPs $j = 1..p$ is a linear sum:

$$GRS(i) = \sum_{j=1}^{p} \beta_j g_{ij} \tag{6.1}$$

GRS can be constructed in a variety of ways [Kooperberg et al., 2010] and have wide application in medicine [Paynter et al., 2010, Yarwood et al., 2015, Abraham et al., 2014]. Given a standard (simplified) logistic regression model for genetic associations with a phenotype $Y \in \{0,1\}$:

$$\text{logit}\{Pr(Y=1)\} = \gamma_0 + \sum_{i=1}^{p} \gamma_j g_{ij} \tag{6.2}$$

where γ_j is the log-odds ratio for the risk of Y given a genotype $g_{\cdot j}$, a natural set of candidate for values of β_j to predict a phenotype Y using a GRS of the form in equation 6.1 is $\beta_j \approx \gamma_j$. In this chapter, I use GRS firstly to differentiate two groups (the EO and PO subtypes of clinical interest) and secondly to differentiate multiple categories (all subtypes).

Assessment of predictive accuracy of a GRS is simple if the values β_i are fitted on a dataset independent of the data on which the GRS is tested. Assessment is more difficult if the data used for fitting the GRS overlap or coincide completely with the data used for testing it. When this was necessary, I used various types of cross-validation. I introduce these as they are used.

6.3 Prediction of extension in oligoarthritic JIA

6.3.1 Motivation

The general aim of this sub-project was to answer the following question:

> Given an individual with newly-diagnosed oligoarthritic JIA, can GWAS data from that individual (or a subset of SNPs derived from GWAS data) be used to

predict, to better accuracy than a random predictor, whether they will progress to polyarthritis?

The motivation for this question is largely clinical rather than scientific, in that a method is sought for differentiation of patients without directly requiring understanding how it does so. Specifically, only a risk score predictive of oligoarthritic extension is sought, rather than explicit identification of EO/PO associated genetic variants. Although the score as a whole may have reasonable predictive power, individual variants contributing to it may not, and may not even be associated with EO/PO status - in the sense that the set of variants with non-zero coefficients in the GRS may have a high FDR.

There is some evidence that extension of osteoarthritis may have some associations with the phenotype at presentation. A retrospective study on 205 American children with JIA found that extension was more likely in children with certain distributions of disease; specifically, wrist or ankle involvement and a symmetric distribution of affected joints [Al-Matar et al., 2002]. Another small study [Gardner-Medwin et al., 2006] suggested that extended oligoarthritis may be characterised by a widespread inflammatory process to a greater extent than persistent oligoarthritis. Evidence of anatomical or physiological variables being predictive of oligoarthritis extension suggest at least some inherent physiological basis to the EO/PO phenotype. If genetic predictors can be found which have independent effects to clinical predictors, these could enable improved clinical predictability of the extension phenotype.

6.3.2 General methods

Penalised regression

A useful set of tools in GRS are penalised regression methods [Tibshirani, 1996], which can simultaneously perform variable selection and fitting, enable use of genomic data with minimal need for an initial dimensionality reduction step. For a (normalised) set of n observations (x_i, y_i), $i \in 1..n$, where each x_i is a vector of p predictors x_{ij}, $j \in 1..p$, the standard ordinary linear model is

$$\begin{aligned} y_i &= x_i^T \beta + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned} \tag{6.3}$$

for which the best linear unbiased estimator for β is

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_i \left(y_i - x_i^T \beta\right)^2 \tag{6.4}$$

By contrast, penalised-regression estimators β include

$$\arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_i \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_j |\beta_j| \right\} \qquad \text{L1-penalised, 'Lasso'}$$

$$\arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_i \left(y_i - x_i^T \beta\right)^2 + \lambda \sum_j |\beta_j|^2 \right\} \qquad \text{L2 - penalised, 'Ridge'}$$

$$\arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_i \left(y_i - x_i^T \beta\right)^2 + \lambda_1 \sum_j |\beta_j| + \lambda_2 \sum_j |\beta_j|^2 \right\} \qquad \text{'Elastic net'}$$

where values $\lambda / \lambda_1, \lambda_2$ are typically chosen to minimise the cross-validated error across the dataset. The 'Lasso' estimator, which is used in this chapter, has the advantage of 'shrinking' unused components of β to 0, imposing sparsity on the model. Lasso estimators are typically strongly consistent if $\lambda = o(p)$ [Chatterjee and Lahiri, 2011].

GRS derived from penalised regression can outperform those fit using an ordinary linear model on variants reaching a pre-fixed threshold on association [Abraham et al., 2014]. From an information-theoretic standpoint, a penalised regression essentially corresponds to a data-driven selection of an optimal association threshold.

Bayesian penalised regression

The maximum-likelihood estimator for β in lasso regression may be considered a posterior mode for β with Laplace (double-exponential) priors on each β_j [Park and Casella, 2008]. This invites the use of variable prior variances across elements of β (known as 'Adaptive lasso' [Zou, 2006]) as a means to lever a lasso procedure on other information:

$$P(\beta) \propto \prod_j e^{\lambda_j |\beta_j|} \tag{6.5}$$

leading to a maximum-likelihood (or equivalently maximum a-posteriori) estimator for β given $X = \{x_{ij}\}$, $y = \{y_i\}$, $\{\lambda_j\}$ and λ_0

$$MLE(\beta | X, y, \{\lambda_j\}, \lambda_0) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_i \left(y_i - x_i^T \beta\right)^2 + \lambda_0 \sum_j \lambda_j |\beta_j| \right\} \tag{6.6}$$

In a logistic regression model, for which $Pr(y_i = 1) = \text{logit}(x_i^T\beta) = (1+\exp(-x_i^T\beta))^{-1}$, the term $\frac{1}{n}\sum_i \left(y_i - x_i^T\beta\right)^2$ is replaced with the standard objective function for logistic regression:

$$l(\beta|\{x_i, y_i\}) = \log\left\{\prod_{i|y_i=1} \text{logit}(x_i^T\beta) \prod_{i|y_i=0} \left(1 - \text{logit}(x_i^T\beta)\right)\right\} \tag{6.7}$$

A high value of λ_j indicates a differential shrinkage favouring non-inclusion in the model, and a low value favours inclusion. The values λ_j are normalised so $\sum_j \lambda_j = p$, and the optimum value of λ_0 (representing the overall shrinkage) is fitted using cross-validation as in the usual Lasso procedure.

Prior distributions which are not conditional on σ^2 may lead to a bimodal posterior for β. However, while this is problematic in the estimation of the posterior distribution by Gibbs sampling or similar [Park and Casella, 2008], it is generally not a serious problem when only seeking the MAP estimate.

Since only the MAP estimate is sought and used, and the posterior distribution is not otherwise considered, this exercise is not a true Bayesian analysis, and the set of values $\{\lambda_j\}$ does not strictly define a prior. However, I consider it helpful to think of distribution 6.5 as a prior-like initial estimate of effect size, and will refer to is as a 'prior distribution' throughout this section, in a slight abuse of terminology.

Leverage strategy

Construction of a predictive GRS for EO/PO is difficult because of the relatively small number of samples and the difficulty of phenotypic ascertainment. In a similar philosophy to the analysis of rare diseases using cFDR in chapter 3, this can be partly alleviated by incorporating data from related phenotypes into the procedure by which coefficients β_j in the GRS are fitted (equation 6.1). As with cFDR, this generally comes at the cost of sacrificing consistency of effect size estimates, due to favouring of variants also associated with the levering trait. Indeed, the estimators of coefficients of variants, and the choice of what coefficients are included in the model, do not represent the true genetic architecture of the EO/PO phenotype. However, in light of the clinical rather than scientific aim, this is not a serious problem.

An important consideration in designing a GRS was how much influence the levering trait should have over the final GRS. A GRS-fitting procedure involves a variable-selection step and a coefficient-fitting step, and summary statistics from a levering phenotype may be used to varying extents in either or both of these. I considered several strategies: firstly, constructing a completely EO/PO agnostic GRS using only the levering phenotype for both

steps (section 6.3.3); secondly, constructing a GRS using only the EO/PO phenotype for both steps with no leverage (section 6.3.4); and thirdly, using a mixture of EO/PO phenotype and the levering phenotype to varying degrees (section 6.3.4).

Choice of dimensionality

Many GRS used in predictive capacities use only confirmed or high-information variants (eg [Paynter et al., 2010]) and although optimal models often include sub-genome-wide associations [Abraham et al., 2014] these generally do not contribute a large proportion of the overall predictive power.

The construction of a GRS using only genome-wide significant variants is an extreme version of a method in which a generalised linear model (usually logistic regression or similar) is used to predict the outcome variable (EO/PO) status using variants reaching some threshold level of significance for association. Choice of the significance threshold for variants to be included in the GRS represents a trade-off with a lower threshold corresponding to a higher proportion of non-associated variants included in the model but a higher proportion of heritability explained.

An ideal compromise is to make a data-informed judgement of the number of predictors to use. For the EO/PO agnostic GRS for T1D and RA, the determination of whether a predictor would be included in the model was entirely on the basis of its p-value for association with the leveing phenotype; that is, all variants up to a maximum p-value were included The absence of available genotypes for RA and T1D precluded more complex variable-selection strategies such as penalised regression.

For EO/PO informed GRS, L-1 penalised regression was used for both variable selection and coefficient fitting, with incorporation of some pre-selection of variables to minimise computational load.

6.3.3 Construction of EO/PO-agnostic GRS

Initially, I constructed GRS fitted to differentiate RA, T1D, and JIA samples from controls. The aim of this procedure was to determine if the pleiotropy between the levering phenotype (RA, T1D, JIA/control status) and the EO/PO phenotype was sufficient that a GRS fitted to the levering phenotype would have predictive ability in the EO/PO phenotype.

GRS for RA

Summary statistics for RA included log-odds ratios (from which I computed p-values p_{RA} against the null hypothesis of no RA association) and effect directions. I identified the set of variants common to the RA and JIA/EO/PO dataset with $p_{RA} \leq 10^{-3}$. I pruned the set of variants by LD to $r^2 < 0.2$, estimating LD between variants using the JIA dataset and prioritising variants with minimum p_{RA}. I then ordered the resulting set of variants V_j by increasing p-value.

Denoting g_{ij} as the numeric genotype (taking values 0,1,2) of individual i at variant j, and $X(V_j)$ as the log-odds ratio of variant j, the GRS value for individual i was defined as

$$GRS_{N(RA)}(i) = \sum_{j=1}^{N(RA)} X(V_j) g_{ij} \tag{6.8}$$

In the absence of genotype data for RA, I determined the optimal value N_{opt} for N using the JIA data:

$$N_{opt} = \arg\max_{N(RA)} \left| t(GRS_{N(RA)}(i) | i \in \text{JIA cases}, GRS_{N(RA)}(i) | i \in \text{Controls}) \right| \tag{6.9}$$

where $t(\cdot,\cdot)$ refers to the t-score from a two-sample t-test against the null hypothesis of equal means of GRS scores in EO and PO populations. Since the EO/PO dataset is independent of the control dataset for JIA, the GRS is unbiased for EO/PO; that is, under the null hypothesis that EO/PO and RA have no shared genetic basis, the GRS has identical expected value in EO and PO populations. The coefficients of the GRS are not agnostic of JIA/control case labels and the GRS is biased in assessments of predictability of JIA/control status on the current dataset; the expected values of the GRS amongst the JIA cases and amongst the JIA controls .

GRS for T1D

Construction of a GRS for T1D presented an additional difficulty as summary statistics were based on score tests and hence effect directions were not available for variants in the dataset. Firstly, absolute log-odds ratios $|X|$ were reconstructed from p-values P and minor-allele frequencies M using the formula

$$|X| \propto \frac{-\Phi^{-1}\left(\frac{P}{2}\right)}{\sqrt{M(1-M)}} \tag{6.10}$$

making use of the approximations (for autosomal SNPs)

$$X \sim N\left(0, \sqrt{M(1-M)}\sqrt{\frac{1}{2n_{control}} + \frac{1}{2n_{case}}}\right) \tag{6.11}$$

$$\left|\Phi^{-1}\left(\frac{P}{2}\right)\right| \sim |N(0,1)| \tag{6.12}$$

under the null hypothesis of no association. The set of variants common to both the T1D and JIA dataset was identified and pruned as for the RA GRS.

The optimal number of predictors and the signs of the coefficients were estimated in parallel from the JIA/control dataset. The optimal number of predictors cannot be determined from the same dataset from which signs are estimated, as a greater number of predictors will always give better discrimination if signs are determined between the groups the GRS seeks to predict. This procedure thus required a cross-validation (XV) procedure, for which ten folds were used.

For each fold, I estimated the signs of coefficients from the nine-tenths of the data not in the fold, and the optimal N was determined on the remaining tenth of the data using a similar formula to equation 6.9 above. The optimal number of predictors was then taken as the median of the ten estimates from the XV folds, and signs for the final GRS were estimated from the whole dataset.

Again, the resulting GRS for T1D would be biased upwards (in favour of better discrimination) for prediction of JIA/control status, but is unbiased for prediction of EO/PO status under the null hypothesis of no shared genetic basis between EO/PO and T1D.

GRS for JIA

Construction of a GRS for JIA could use a wider range of methodologies due to the availability of genotypic data. I used two separate methods, generating two separate sets of predictors. Model parameters were again fitted using cross-validation.

The first GRS was fitted in a similar way to the GRS for RA. I ranked variants by decreasing absolute log-odds ratio between JIA and control samples and pruned by LD to $r^2 < 0.2$. The GRS was computed with the same form as equation 6.8. The optimal number of predictors was determined by producing, ranking and pruning ten sets of p-values p^k_{EO}, $k \in 1..10$, each with one-tenth of the data removed, and fitting a GRS of the form of 6.8 to each set of statistics.

I then tested each fold-specific GRS on the remaining one-tenth of the data, and determined the number of predictors which maximally separated subgroups in a similar way to equation 6.9. The final number of predictors was chosen as the median of these ten values.

I fitted the second GRS to the pruned set of variants using a lasso model (equation 6.6) with equal prior variances $\lambda_j \equiv 1$. I determined an optimal value of λ_0 by cross-validation, using the same ten-fold subdivision of the data as for the previous GRS.

Again, both GRS were overfitted to JIA and measures of separation would be biased upwards if the GRSs were used to predict JIA/control status on the same set of data used to generate the GRS.

Assessment of predictive accuracy

For all EO/PO agnostic GRS, a sufficiently large number of variants were included in the final model that GRS scores could be assumed to be normally distributed within each JIA subtype. Under this assumption, performance was compared using a standard two-groups t-test, to test for difference in mean GRS between EO and PO individuals.

Predictive performance was also visually assessed using receiver-operator characteristic (ROC) curves, and estimation of sensitivity and specificity at an optimal threshold determined by maximal Youden's index (sens. + spec. -1).

6.3.4 Construction of EO/PO informed GRS

The principal idea of this approach was to use a large dataset on a related disease (JIA, RA, T1D) to set the values λ_j used in the Laplace prior (equation 6.5), followed by GRS construction using Lasso regression on the EO/PO samples.

Variable selection

The time-complexity of the LARS algorithm used for computing Lasso estimators with p predictors on n samples is $O(p^3 + p^2 n)$ [Efron et al., 2004], which becomes very computationally intensive for $p > 1000$, and for whole-genome scale estimation in which $p >> 10^5$ the LARS algorithm is not practically applicable. Under the assumption that the proportion of non-associated SNPs π_0 is ≈ 1 (with the standard two-groups model of [Efron et al., 2008]), some feature selection is justified before application of the LARS algorithm.

I did this by restricting to variants showing moderate evidence of association with the levering phenotype or with the EO/PO phenotype ($p < \alpha_T$; I used $\alpha_T = 1 \times 10^{-3}$). This required using the ten sets of summary statistics p^k_{EO}, $k \in 1,..10$ used in fitting the GRS for

JIA in the previous section, in which the set p^k_{EO} compares EO and PO samples with samples in fold k removed from the analysis. Variable selection was performed separately for each XV fold k, assessing EO/PO association using summary statistics p^k_{EO}. In order that the GRS fitted to each nine-tenths of the data to be completely agnostic to the remaining one-tenth, the cross-validation procedure to fit the value λ_0 (see equation 6.6) was nested within each fold of the wider cross-validation procedure.

Overall EO/PO informed GRS algorithm

The overall algorithm for fitting and assessing the GRS was as follows

Definitions

S: genome-wide set of SNPs
T^1, T^0: case/control cohorts for trait under investigation (EO/PO), genotyped at S
$X_T(P,Q)$: genotype submatrix for trait under investigation at samples P, variables Q.
$y_T(P)$: case/control indicator for trait under investigation at samples P
p_L: p-values for trait used for leverage, genotyped at S
$p_{A,B}$: set of p-values from comparing cohort A to cohort B
α_L: cutoff p-value in trait used for leverage for SNP inclusion in lasso model
α_T: cutoff p-value in trait under investigation for SNP inclusion in lasso model
$prune(S',X^1,X^0)$: function pruning set of SNPs S' by LD, favouring high p_{X^1,X^0}
Λ: set of potential values for λ_0 with kth element Λ_k
$f(.)$: function mapping p-values p_L to prior parameters $\{\lambda_j\}$ for lasso model
$MLE(\beta|X,y,\{\lambda_j\},\lambda_0)$: MLE for β under logistic formulation of equation 6.6
$m(y_1,y_2)$: metric for accuracy of GRS values y_1 for predicting phenotype y_2

function GENERATE UNBIASED GRS

$N \leftarrow 10$ ▷ Number of XV-folds
$T^0_1, T^0_2, \ldots T^0_N \leftarrow$ partitioning of T^0 for XV
$T^1_1, T^1_2, \ldots T^1_N \leftarrow$ partitioning of T^1 for XV
$S_L \leftarrow \{SNPs : p_L < \alpha_L\}$ ▷ SNPs associated with levering trait
for $k \in 1...N$ **do** ▷ Find sets of SNPs to include in each XV fold
$T^1_A \leftarrow \bigcup_{s\neq k} T^1_s$, $T^0_A \leftarrow \bigcup_{s\neq k} T^0_s$ ▷ Training sets
$T^1_B \leftarrow T^1_k$, $T^0_B \leftarrow T^0_k$ ▷ Test sets
$S^k_T \leftarrow \{SNPs : p_{T^1_A,T^0_A} < \alpha_T\}$ ▷ SNPs associated with EO/PO, fold-k agnostic
$S^k \leftarrow S^k_T \cup S_L$ ▷ SNPs to be included in lasso model, fold-specific

$S^k \leftarrow prune(S^k, T^1, T^0)$

$\{\lambda_k\} = f(p_{T_A^1, T_A^0}(S^k))$ ▷ Prior parameters for SNPs in S^i

end for

initialise $Y : |Y| = |T^0 \cup T^1|$ ▷ Y will be set to overall unbiased GRS

for $k \in 1..N$ **do**

define $T_A^1, T_A^0, T_B^1, T_B^0$ as above

initialise $Y_k : dim(Y_k) = (|T_A^1 \cup T_A^0| \times |\Lambda|)$ ▷ GRS for samples not in fold k

for $h \in \{1..N\} \setminus k$ **do**

$T_C^1 \leftarrow \bigcup_{s \neq k,h} T_s^1$, $T_C^0 \leftarrow \bigcup_{s \neq k,h} T_s^0$

$T_D^1 \leftarrow T_h^1$, $T_D^1 \leftarrow T_h^0$

for $l \in \Lambda$ **do**

$\beta_{khl} \leftarrow MLE(\beta | X_T(T_C^1 \cup T_C^0, S^h), y_T(T_C^1 \cup T_C^0), \{\lambda_j\}, \Lambda_l)$

$Y_k(T_D^1 \cup T_D^0, l) = X_T(T_D^1 \cup T_D^0, S^h)\beta_{khl}$ ▷ GRS in fold k, $\lambda_0 = \Lambda^l$

end for

end for

$\lambda_0^k = \{\Lambda_l : m(Y_k[\cdot, l], y_T(T_A^1 \cup T_A^0))$ is maximised$\}$ ▷ Find best λ_0 from XV

$\beta_k \leftarrow MLE(\beta | X_T(T_A^1 \cup T_A^0, S^k), y_T(T_A^1 \cup T_A^0), \{\lambda_j\}, \lambda_0^k)$ ▷ Fold-k agnostic

$Y(T_B^1 \cup T_B^0) = X_T(T_B^1 \cup T_B^0, S^k)\beta_k$ ▷ Unbiased GRS values for samples in fold k

end for

return $Y, m(Y, y_T)$

end function

□

Since all the inputs into generating the GRS for samples in fold k - $X_T(T_A^1 \cup T_A^0, S^k), y_T(T_X^1 \cup T_X^0), \{\lambda_k\}$, and λ_0^k - are agnostic of phenotype labels y_T from fold k, the GRS has equal expected values across individuals with $y_T = 1$ and individuals with $y_T = 0$ under the null hypothesis that y_T is independent of all genetic predictors.

Choice of function f

There are several choices of the function f in the previous section, which determines the relationship between an observed p-value p_i in the trait used for leverage and the parameters λ_j (corresponding to variance $2/\lambda_j^2$) of the Laplace prior (equation 6.5) used in fitting the lasso model. Since the value of λ in equation 6.6 is variable and fitted by cross-validation, the set of values $\{\lambda_j\} = f(\{p_j\})$ may be multiplied by an arbitrary constant. All functions f

were thus scaled such that

$$\sum f(p_j) = \sum \lambda_j = |\{\lambda_j\}| = p \tag{6.13}$$

where p is the total number of predictors in the model. I proposed three prior-generating functions f_1, f_2, f_3 and fitted a separate GRS using each of them.

The first and simplest form of f is a theshold function:

$$f_1(p_j) \propto \begin{cases} x & \text{if} \quad p_j < \alpha \\ y & \text{if} \quad p_j \geq \alpha \end{cases} \tag{6.14}$$

that is, modulating variances according to whether variants reach a p-value threshold α in the levering trait. This approach is used implicitly with $x = 1$, $y = \infty$ in the pruning step in the algorithm. A disadvantage is that the choice of α, x, y are somewhat arbitrary. A prior of this form was used with $x = \frac{1}{2}$, $y = 1$ with the value $\alpha = 5 \times 10^{-8}$ corresponding to the standard genome-wide significance threshold. I chose these specific values of (α, x, y) as roughly 'simple' values before fitting any models.

A preferential approach is to choose the parameters λ_j to match the Laplace distributions to a presumed underlying sampling distribution for each observed unsigned Z score $z_j = -\Phi^{-1}(p_j/2))$. If $|z_j|$ is an absolute value of a single observation from a distribution $N(0, \sigma_j^2)$, the MLE for the standard deviation σ_j is equal to $|z_j|$. I sought to match the distribution from which z_j is observed to the Laplace prior, and used the Kullback-Liebler divergence (an asymmetric measure of difference between probability distributions) to do this. The K-L divergence of a Laplace distribution with PDF $l(z; \lambda_j) = \frac{\lambda_j}{2} e^{-\lambda_j |z|}$ and a normal distribution with PDF $n(z; \sigma_j) = (\sqrt{2\pi}\sigma_j)^{-1} e^{-z^2/(2\sigma_j^2)}$ is

$$\begin{aligned} D_{KL}(l||n) &= \int_{-\infty}^{\infty} \frac{\lambda_j}{2} e^{-\lambda_j |z|} \log \left(\frac{\frac{\lambda_j}{2} e^{-\lambda_j |z|}}{\frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{z^2}{2\sigma_j^2}}} \right) dz \\ &= 2 \int_0^{\infty} \lambda_j \log(\sqrt{2\pi}\lambda_j \sigma_j) dz + 2 \int_0^{\infty} \left(-\lambda_j z + \frac{z^2}{2\sigma_j^2} \right) \lambda_j e^{-\lambda_j z} dz \\ &= \log(\lambda_j \sigma_j) + \frac{1}{\lambda_j^2 \sigma_j^2} + \log\left(\sqrt{\frac{\pi}{2}} \right) - 1 \end{aligned} \tag{6.15}$$

which is minimised when $\lambda_j \sigma_j = \sqrt{2}$. Thus the parameter λ_j leading to the closest-matched Laplace distribution $l(z;\lambda_j)$ to the presumed underlying distribution $N(0,\sigma_j^2)$ leads to the second form for f

$$f_2(p_j) \propto \frac{\sqrt{2}}{|z_j|} \propto -\frac{1}{\Phi^{-1}(p_j/2)} \tag{6.16}$$

A disadvantage of this approach is that the prior variance is unbounded, in that SNPs with exceptionally strong associations with the levering trait can be effectively guaranteed inclusion in the lasso model. Since the levering trait is not the same as the trait under investigation, this may weaken the predictive power.

A better quantity on which to base the prior variance may be the posterior probability of association with the trait used for leverage, assuming a two-groups model [Efron et al., 2008]. Assume Z-scores for the levering trait have the distribution

$$Z \sim \begin{cases} N(0,1) & \text{with prob. } \pi_0 \\ N(0,\sigma^2) & \text{with prob. } 1-\pi_0 \end{cases} \tag{6.17}$$

where values π_0, σ may be set using an E-M algorithm (see chapter 2, section 2.4.3), noting that σ has a different meaning to the SNP-specific σ_j used in equation 6.15. The posterior probability ψ_j of association (with regard to the null hypothesis H_0^L for the levering trait) for SNP j with z-score z_j is

$$\psi_j = Pr(H_0^L|z_j) = \frac{Pr(H_0^L)p(z_j|H_0^L)}{p(z_j)} = \frac{\pi_0 N_1(z_j)}{\pi_0 N_1 + (1-\pi_0)N_{\sigma^2}(z_j)} \tag{6.18}$$

where $N_{\sigma^2}(z)$ denotes the PDF of $N(0,\sigma^2)$ at z. The observed z_j can be considered a sample from a distribution of the form 6.17 with π_j in the place of π_0:

$$z_j \sim \begin{cases} N(0,1) & \text{with prob. } \psi_j \\ N(0,\sigma^2) & \text{with prob. } 1-\psi_j \end{cases} \tag{6.19}$$

This distribution has the advantage of favouring inclusion for large z_j, but with bounded expectation equal to the overall variance σ^2 of the effect size distribution across the levering trait as $\psi_j \to 0$.

The value of λ_j minimising the K-L divergence between a Laplace distribution with PDF $l(z;\lambda_j) = \frac{\lambda_j}{2} e^{-\lambda_j|z|}$ and a distribution $s(z;\psi_j)$ of the form of equation 6.19 is independent of ψ_j; considering distribution $s(z;\psi_j)$ as a limit as $a \to 0$ of continuous distributions $s_a(z;\psi_j)$

with PDFs

$$s_a(z;\psi_j) = (1-\psi_j)N_{\sigma^2}(z) + \psi_j \frac{\psi_j}{2a}\mathbf{1}_{|\mathbf{z}|<\mathbf{a}} \tag{6.20}$$

we have

$$\begin{aligned} D_{KL}(l||s) &= \lim_{a\to 0}\left(\int_{-\infty}^{\infty} l(z;\lambda_j)\log\{l(z;\lambda_j)\}dz - \int_{-\infty}^{\infty} l(z;\lambda_j)\log\{s_a(z;\psi_j)\}dz\right) \\ &= -1 - 2(1-\psi_j)\lim_{a\to 0}\left(\int_0^a \frac{\lambda_j}{2}e^{-\lambda_j z}\left(N_{\sigma^2}(z) + \frac{\psi_j^2}{2a(1-\psi_j)}\right)\right. \\ &\qquad \left. + \int_a^{\infty}\frac{\lambda_j}{2}e^{-\lambda_j z}N_{\sigma^2}(z)\right) \\ &= (1-\psi_j)\left(\log(\lambda_j\sigma) + \frac{1}{\lambda_j^2\sigma^2} + \log\left(\sqrt{\frac{\pi}{2}}\right)\right) - 1 \end{aligned} \tag{6.21}$$

which is minimised for $\lambda_j\sigma = \sqrt{2}$ regardless of ψ_j.

In order to choose a Laplace distribution $l(z;\lambda_j)$ corresponding to the distribution $s(z;\psi_j)$, the value λ_j may be fitted on the basis of expectation of absolute values of random variables between $l(z;\lambda_j)$ and $s(z;\psi_j)$, accounting for a 'baseline' possibility of association with the levering phenotype when $\psi_j \approx 1$.

The expected value of $|X|$ if X has Laplace PDF $l(z;\lambda_i) = \frac{\lambda_i}{2}e^{-\lambda_i|z|}$ is $E_l(|X|) = \frac{1}{\lambda_i}$. If X has normal PDF $n(z;\sigma) = N_{\sigma^2}(z)$ then $E_n(|X|) = \sqrt{\frac{2}{\pi}}\sigma$, and if X has distribution $s(z;\psi_j)$ of the form in equation 6.19 then $E_s(|X|) = (1-\psi_j)\sqrt{\frac{2}{\pi}}\sigma$. The third definition of f

$$f_3(p_i) \propto \frac{1}{1+(1-\psi_i)\sqrt{\frac{2}{\pi}}\sigma} \tag{6.22}$$

sets $E_l(|X|) = 1 + E_s(|X|)$.

Choice of function m and assessment of significance

The function $m(Y, y_T)$ measured deviance of the GRS from the true phenotype. Given the potential for different behaviours of y_T in different XV-folds, I fitted a logistic model

$$\text{logit}^{-1}(Pr(Y=1)) = \sum \gamma_k \mathbf{1}_{\text{fold}=k} + \gamma_0 y_T \tag{6.23}$$

and defined $m(Y, y_T)$ as the p-value from a score test against the null hypothesis $\gamma_0 = 0$.

Although the value $m(Y, y_T)$ is derived from a p-value for an observed value γ_{obs} of γ, the value $m(Y, y_T)$ is *not* equal to $Pr(|\gamma| > |\gamma_{obs}| \Big| H_0^{EO/PO})$, where $H_0^{EO/PO}$ is the null hypothesis that the EO/PO phenotype is independent of all genetic variants. Under $H_0^{EO/PO}$, the values y_T and Y are independent for individuals in fold i, since the GRS is agnostic to those samples. However, fold-specific measures of predictive accuracy are not independent between folds, and hence values $m(Y, y_T)$ do not have a $U(0,1)$ distribution under $H_0^{EO/PO}$, despite being derived from p-values.

In order to estimate a true p-value associated with an observed value $m(Y, y_T)$, the entire GRS-generating procedure was repeated > 500 times with permuted phenotype labels, with the final p-value estimated using the quantile of the observed $m(Y, y_T)$ in the distribution of $m(Y, y_T)$ values from permuted phenotypes.

6.3.5 Genetic associations with EO/PO status

Initially, I sought to establish whether the dataset supported any genetic differences between EO and PO samples at all. I performed a GWAS comparing EO- and PO- subgroups of JIA using sex and the three principal components as covariates, generating p-values p_{EO}. Sex does not show evidence of association with extension [Al-Matar et al., 2002] but it is associated both with JIA and oligoarthritis in general [Ravelli and Martini, 2007] suggesting sex-specific mechanisms of oligoarthritis. These may be better-detected with sex adjustment, and the inclusion of sex in the model has little effect on the power to detect sex-independent EO/PO associations. Principal components were included in the model to account for potential regional differences in diagnosis (for example, due to differential follow-up times).

The inflation factor λ for the set p_{EO} was 1.02 ($\lambda_{1000} = 1.03$) indicating a low degree of inflation as expected. No p-values reached genome-wide significance or Bonferroni-corrected significance at $\alpha = 0.05$ (minimum p-value 5.4×10^{-8}; Bonferroni-corrected significance cutoff 6.6×10^{-9}).

A Q-Q plot demonstrated no evidence of departure from $p_{EO} \sim U(0,1)$ (figure 6.1). Collectively, isolated analyses of the EO/PO phenotype suggest a total absence of discoverable heritability. However, GWAS Q-Q plots can also show no evidence of association for underpowered studies (which this is likely to be), whilst larger studies on the same disease show extensive evidence of associations.

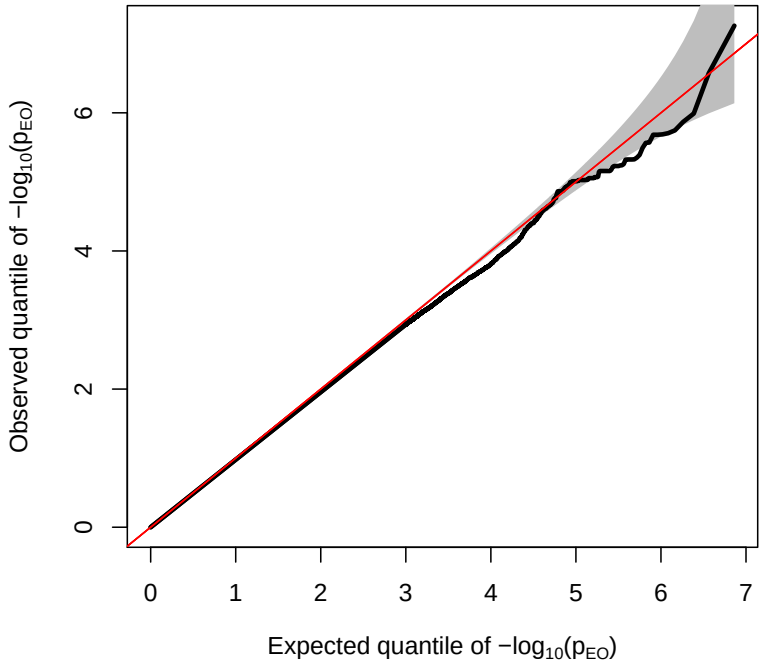


Fig. 6.1 Q-Q plot for p-values derived from comparison between extended and persistent oligoarthritis. The grey region shows a pointwise 99% confidence envelope for each order statistic of a uniform distribution. There is no evidence of departure from a $p_{EO} \sim U(0,1)$.

Table 6.3 Details of GRSs fitted to T1D, RA or JIA for predicting EO/PO phenotype. Column LP stands for levering phenotype. Column $dim(GRS)$ is the number of variables in the GRS.

LP	$dim(GRS)$	**P-value**
RA	69	2.1×10^{-4}
T1D	1113	3.2×10^{-4}
JIA 1 (log-OR)	91	3.8×10^{-4}
JIA 2 (lasso)	42	2.0×10^{-5}

Conditional associations

Despite no obvious departure of p_{EO} from $U(0,1)$, I examined whether conditioning on RA, T1D or JIA association revealed any inflation. As discussed in chapter 5, association statistics from JIA/control and EO/PO are independent under the null hypothesis for EO/PO. Figure 6.2 shows the results of this analysis. Notably, there is visible inflation in all cases when the MHC region (chr6:25-40Mb, NCBI build 37) is included, which disappears when the MHC region is removed, indicating that visible inflation is generally driven by the MHC region. Given the extensive involvement of MHC in JIA and autoimmune disease in general [The Wellcome Trust Case Control Consortium, 2007, Hinks et al., 2013] this is reasonable evidence for some genetic basis to oligoarthritic extension, principally driven by the MHC region.

Despite the visual evidence, analyses using cFDR found no EO/PO associations when conditioning on any of the three phenotypes, even using liberal thresholds $\alpha_1(EO/PO|\cdot)$ (see chapter 3, section 3.2.2) for genome-wide association.

6.3.6 GRS results

EO/PO agnostic GRS

All four EO/PO agnostic GRS (fitted to RA, fitted to T1D, fitted to JIA using odds-ratios and fitted to JIA using lasso) could significantly discriminate EO and PO samples. Results from the comparisons are shown in table 6.3. The best-performing GRS was the lasso-based model fitted to JIA data. Figure 6.4 shows the respective densities of the GRS fitted to JIA using the lasso model in EO and PO samples.

ROC curves for the four predictors showed modest predictability. Sensitivity and specificity of predictors are shown in table 6.4. These results indicated the presence of pleiotropy between each of RA, T1D, and JIA and the EO/PO phenotype, to the extent that EO/PO is partially predictable on the basis of genetic susceptibility to RA, T1D, or JIA alone.

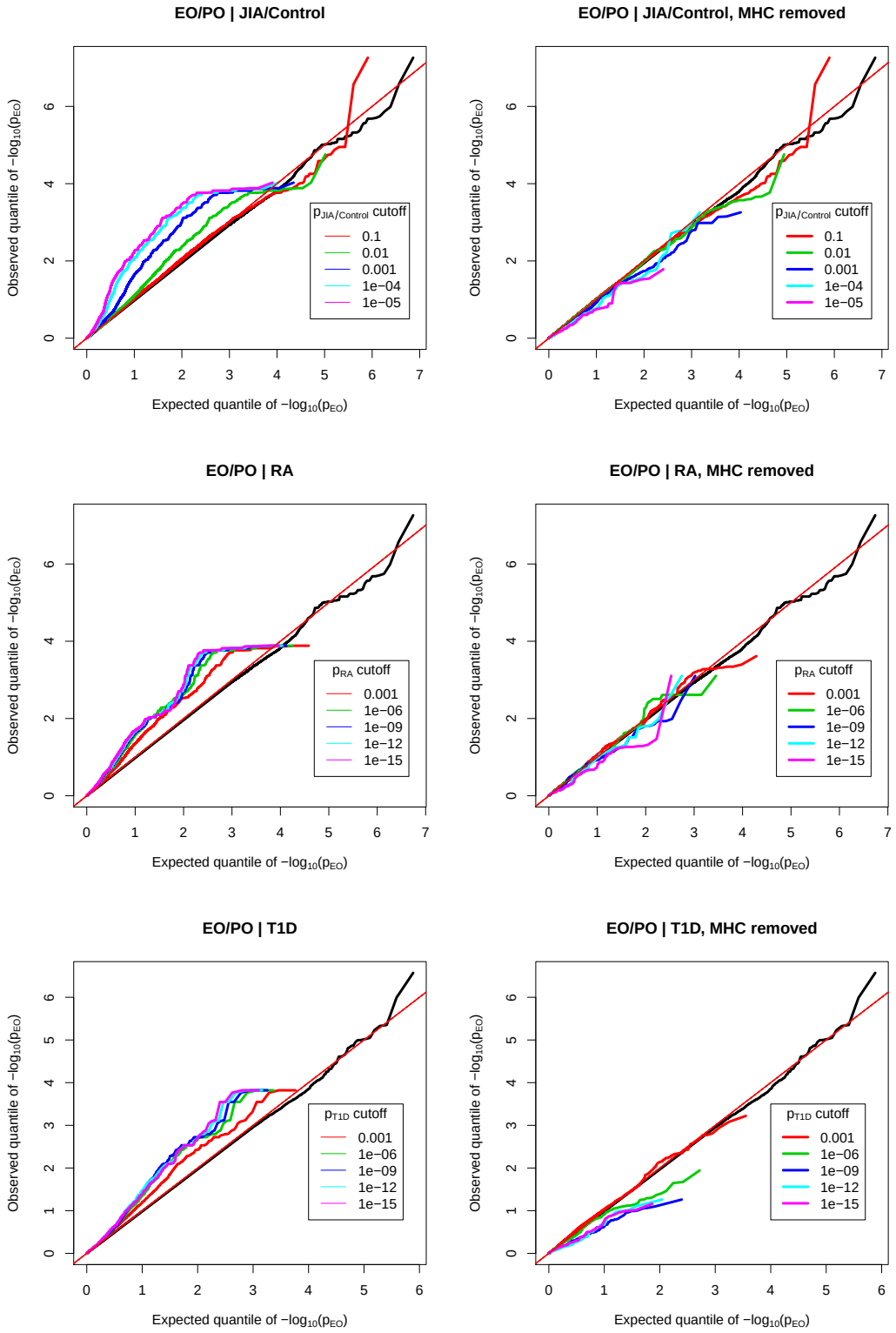


Fig. 6.2 Conditional Q-Q plot for p-values derived from comparison between extended and persistent oligoarthritis, conditioning on JIA, T1D and RA, with MHC included (left panels) and excluded (right panels). Note different thresholds for conditional p-values, chosen because the studies on RA and T1D were larger than the study on JIA. There is evidence of deviation from a uniform distribution for p_{EO} when conditioning on association with other phenotypes. Confidence envelopes are not shown as they differ for each conditional p-value threshold. The deviation in distribution disappears when the MHC region is removed from the analysis, suggesting that the inflation is MHC driven.

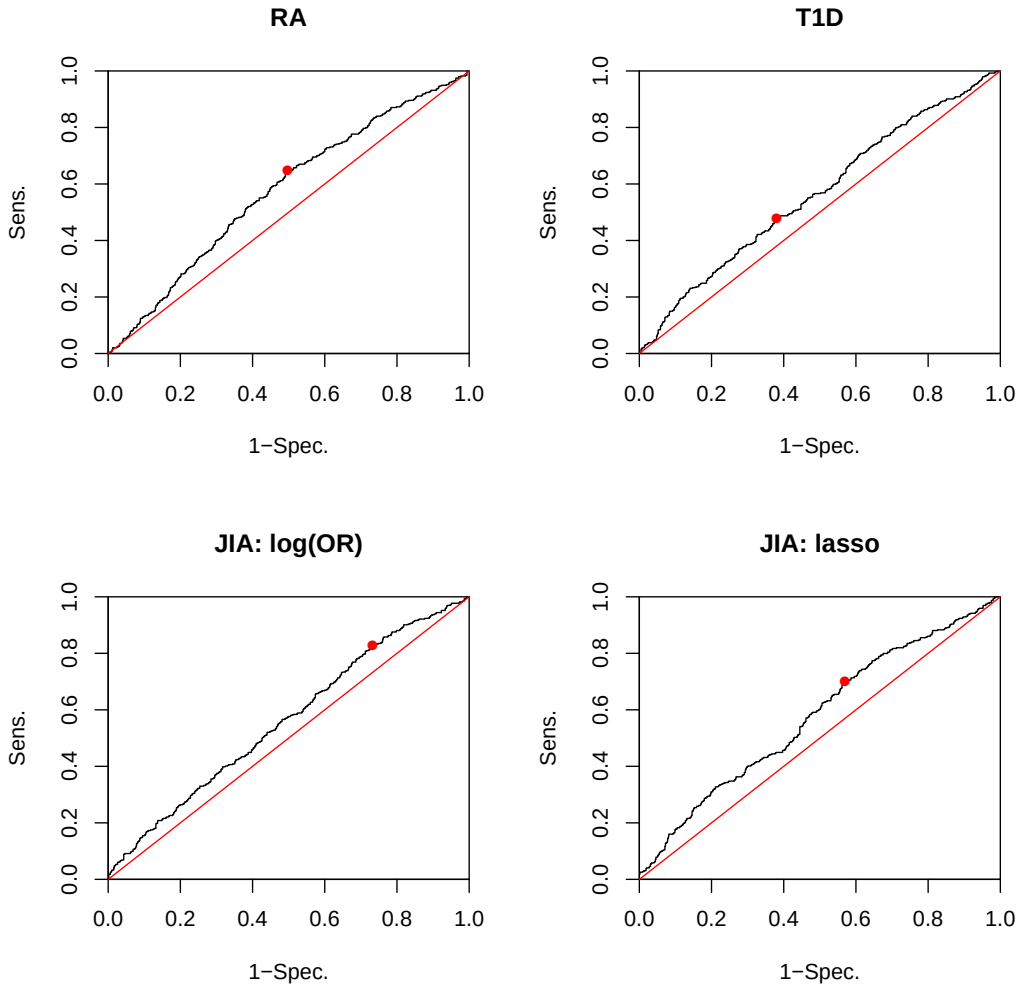


Fig. 6.3 ROC curves for predicting EO/PO using GRS for T1D, RA and JIA. Modest predictive power is seen in each case. Red dots mark the point of maximal Youden's index (sens. + spec. -1).

Table 6.4 Sensitivity and specificity at maximal Youden's index of GRSs fitted to T1D, RA or JIA for predicting EO/PO phenotype.

LP	AUC	Sens. (%)	Spec. (%)
RA	0.57	65	50
T1D	0.56	48	62
JIA 1 (log-OR)	0.56	83	27
JIA 2 (lasso)	0.58	70	43

EO/PO informed GRS

After pruning to only SNPs with $p^k_{EO} < \alpha_T = 1 \times 10^{-3}$ and pruning by LD (in the same way as described in chapter 2), a median of 1345 (range 1316-1384) SNPs were included in the lasso model for each XV fold. A median of 3 (1-38) variants were included in the final predictive score. The final predictive scores were essentially identical whether or not SNPs associated with each of the levering trait were included in the lasso model, although with different fitted values of λ due to different numbers of input variables. Results from the GRS analysis are shown in figure 6.5. The unbiased GRS with no leverage could not discriminate between groups ($p = 0.36$)

The GRS with leverage in general performed much better, in that more variants were included in the model, and there was more discrimination between EO and PO by the unbiased GRS (table 6.5), most strikingly when levering on T1D. Only one comparison (levering on T1D, with prior f_3) reached Bonferroni-corrected significance ($p < 0.05/10 = 0.005$; ten comparisons shown in table 6.5) although several (levered on JIA with prior f_1, levered on JIA with prior f_3, levered on T1D with prior f_2, levered on T1D with prior f_3) showed suggestive associations at FDR≤ 0.1 ($p < 0.033$).

None of f_1, f_2, f_3 were clearly dominant, although f_2 appeared the weakest. A graph of densities from a successful GRS is shown in figure 6.6. As an illustration of predictive power, an ROC curve is shown for the GRS levered on T1D with prior f_1 (figure 6.7). With this predictor, the cutoff with optimal Youden's index (sens. + spec. -1) had sensitivity 69% and specificity 45%. These values are likely to be overestimates of the true performance of this GRS on new data, since it was chosen as the best-performing GRS out of those tested.

6.3.7 Discussion

In this project, I determined that the EO/PO phenotype has a genetic component, and that it may be predicted to a better-than-random degree using GRS. This indicates that genetic testing of patients with newly-diagnosed oligoarthritis could provide a clinically useful classification (if it provides independent predictive ability to clinical variables), ultimately allowing more efficient management of patients with JIA. I also developed a generic method for generating and testing a GRS leveraged on data from a different, related phenotype, using standard machine-learning procedures.

This exercise serves as at best a pilot study, and is not on its own enough to suggest changes to patient management protocols. Sample sizes were too small to identify any individual variant associated with the EO/PO phenotype, and no independent dataset was

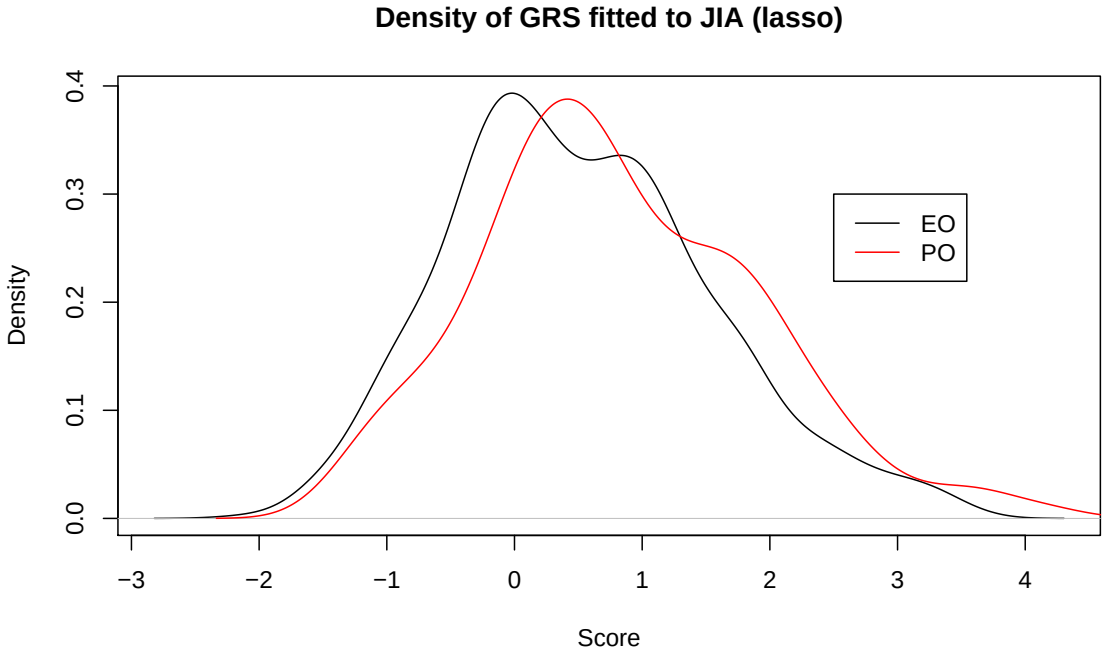


Fig. 6.4 Results of GRS fitted to the JIA/control phenotype for predicting extension in oligoarthritis. Significant discrimination is evident ($p = 2.0 \times 10^{-5}$).

Table 6.5 Details of GRSs fitted to EO/PO phenotype. Column LP stands for levering phenotype. Columns $dim(model)$ and $dim(GRS)$ give the number of variables entered into the lasso model and the number of variables in the GRS respectively. Columns $dim(model)$, and $dim(GRS)$ are medians across XV-folds.

LP	**Prior**	$dim(model)$	$dim(GRS)$	**P-value**
None	None	1346 (1316-1384)	3 (0-38)	0.36
JIA	f_1	3362 (3335-3407)	13 (9-38)	0.015
JIA	f_2	"	29 (0-125)	0.080
JIA	f_3	"	35 (0-74)	0.030
RA	f_1	3060 (3031-3104)	13 (1-51)	0.75
RA	f_2	"	6 (4,10)	0.14
RA	f_3	"	0 (0,23)	0.20
T1D	f_1	2633 (2608-2674)	19 (10-51)	< 0.001
T1D	f_2	"	7 (5-10)	0.027
T1D	f_3	"	15 (4-39)	0.015

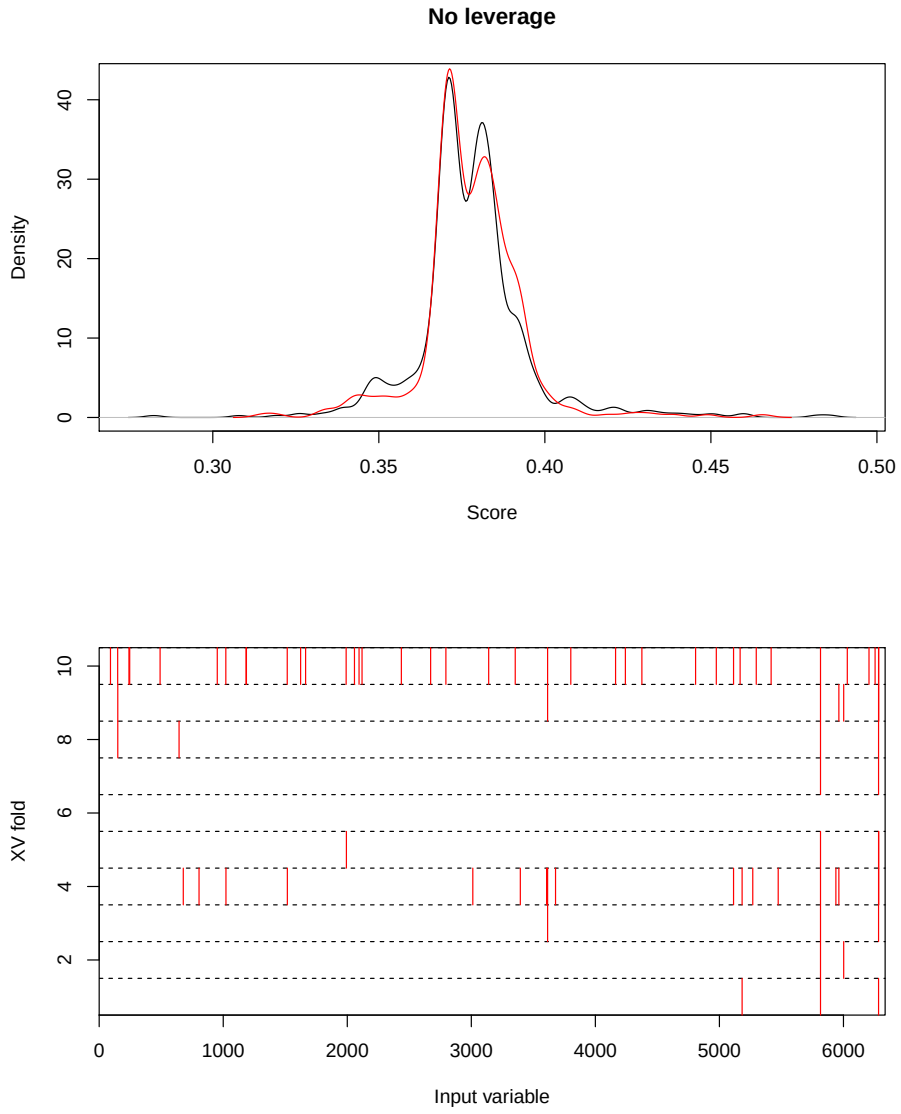


Fig. 6.5 Results of GRS fitted to EO/PO phenotype with no leverage. Top panel shows 'unbiased' GRS scores for EO and PO samples. No discrimination is evident ($p = 0.36$). The lower panel shows variable inclusion in the GRS, with a red line indicating a nonzero coefficient. There is reasonable concordance in the variables included in the models for each cross-validation fold.

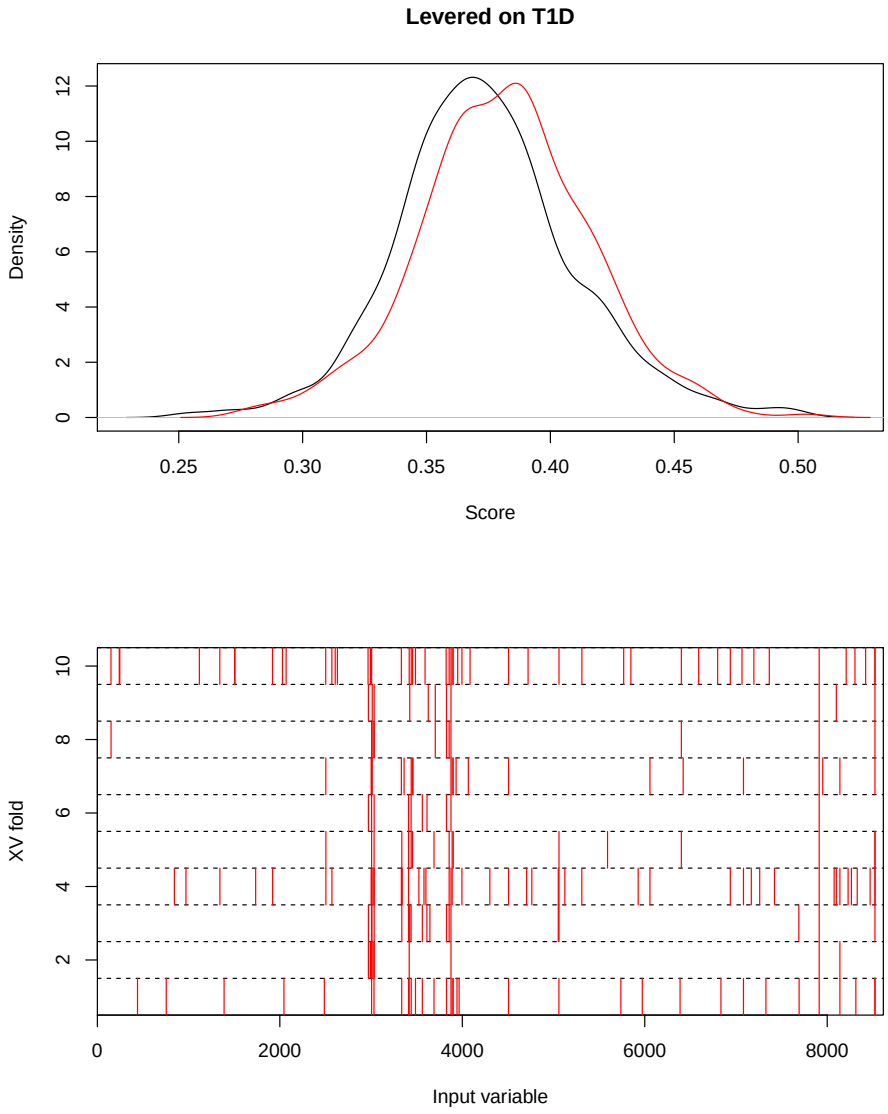


Fig. 6.6 Results of GRS fitted to EO/PO phenotype leveraged on T1D using prior form f_1, as an example of a successful model. Top panel shows 'unbiased' GRS scores for EO and PO samples. Moderate discrimination is evident ($p = 3.0 \times 10^{-4}$). The lower panel shows variable inclusion in the GRS, with a red line indicating a nonzero coefficient. There is good concordance in the variables included in the models for each cross-validation fold.

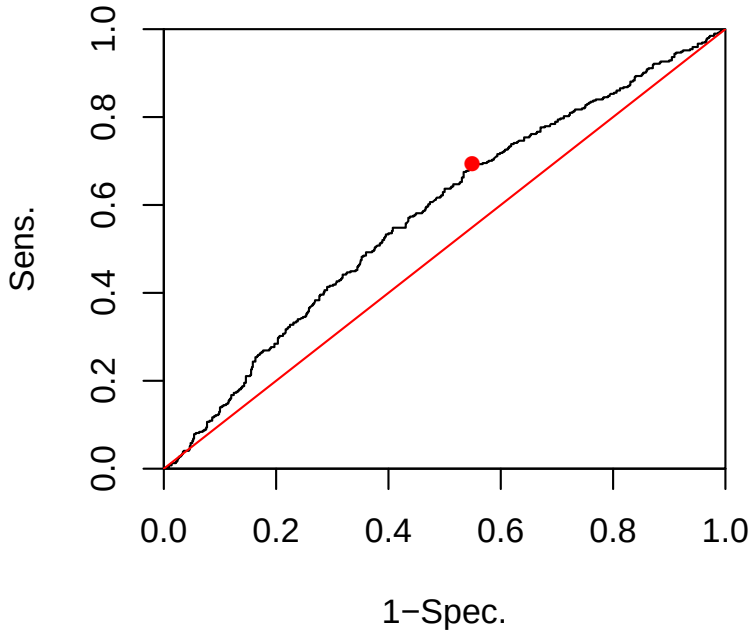


Fig. 6.7 ROC curve for GRS on EO/PO phenotype, levered on T1D with prior f_3. At the point of maximal Youden's index (sens. + spec. -1), the classifier has sensitivity 69% and specificity 45%. These values should not be interpreted as estimations of performance of the GRS on new data, since this model was chosen as the best-fitting example.

available for validation of the proposed risk scores. The current study has no way to determine if between-group discrimination arises from the EO/PO phenotype or a confounder.

Generalisability and confounders

The generalisability of the GRS is dependent on the GWAS samples' representation of of the true population presenting with oligoarthritic JIA. This cohort has the advantage that patients were recruited agnostic of EO/PO phenotype, on the basis of presentation with oligoarthritis only. This reduces the chance that the observed genetic differences were due to confounding variables. However, the presence of confounding remains a possibility. One potential such confounder is misdiagnosis. If some number of samples had only transient arthralgia or oligoarthritis secondary to some other disease process, we may expect that such samples should be over-represented in the PO group, since the arthritis would not progress. This could mean that genetic differences between EO and PO groups were due only to the higher proportion of 'true' JIA cases in the former. This would be consistent with the observed inflation of EO/PO conditioning on JIA association (figure 6.2).

It may be argued, however, that confounding of this type does not matter from a clinical perspective. The population on whom such a classifier may be used matches that in the study - that is, children presenting with apparent oligoarthritis - and will contain misdiagnosed cases at a similar rate to the study cohort. In this context, the clinical usefulness of predicting extension does not depend on what the prediction is based on. The same is true if the cohorts are confounded by ethnicity or geographic location, although in this case the GRS would not be independent of clinical or epidemiological predictors. Confounders are only problematic in this application if they are unique to the current dataset, and such confounders should be eliminated by the sampling procedure.

GRS performance

This exercise broadly used three types of predictive GRS: firstly, GRS fit entirely on the basis of external datasets on related diseases fitting (sections 6.3.3, 6.3.3, 6.3.3); secondly, GRS which used EO/PO data in combination with summary statistics from related diseases (table 6.5, rows 2-10), and thirdly, GRS fitted only on EO/PO genotypes (table 6.5, row 1). Comparisons of predictive accuracy between the first type of GRS and the second two must be made cautiously, as the second two required use of the same dataset for both training and testing. However, it appeared that both the first and second types of GRS performed better than the third.

This would not be expected to hold as sample sizes increase to infinity, as the genetic architectures of RA, T1D and JIA are at best approximations of the genetic architecture of EO/PO. Indeed, the coefficients of the third type of GRS should be consistent estimators of variant effect sizes in EO/PO, the coefficients of the first type should be consistent estimators of effect sizes in the levering phenotype, and the coefficients of the second type should converge to some average of the two (assuming that the sample sizes in EO/PO and the levering phenotype increase at approximately the same rate). This should result in a reversal of the observed predictive accuracies, with the third type of GRS ultimately performing the best, followed by the second, followed by the first.

The observed results, in conjunction with the conditional analyses in section 6.3.5, do indicate a degree of pleiotropy between the genetic basis of EO/PO and the three levering phenotypes, principally at the MHC locus. Notably, this only demonstrates pleiotropy in the sense of sharing of associated, rather than causal variants. All of the GRS use implicit effect size estimates for EO/PO for prediction, and in the current study, the improved performance of EO/PO agnostic GRS suggest that the systematic error in these estimates in the GRSs for T1D/RA/JIA due to the difference in phenotype is less than the random error in the estimates from the small sample size in the EO/PO phenotype itself.

The relatively poor performance of the GRS levered on RA compared to those levered on T1D and JIA was surprising, given the performance of the GRS for RA and the evidence of pleiotropy with EO/PO (figure 6.2). The size and precision of the GWAS may have been a problem; considerably more variants reached $p_{RA} < 5 \times 10^{-8}$ than did $p_{T1D} < 5 \times 10^{-8}$ or $p_{JIA} < 5 \times 10^{-8}$, presumably including many which were specifically associated with RA. Variants with $p_{RA} < 5 \times 10^{-8}$ were given the same weight in prior f_1 and approximately the same prior weight in f_3, and RA-specific variants may have diluted the effect of variants which were associated with both EO/PO and RA.

Evaluation of method

The choice of methods was largely due to necessity. Penalised regression is empirically the strongest form of GRS in typical complex-disease settings [Abraham et al., 2013, Kooperberg et al., 2010]. Pre-selection of variables on the basis of p-value (within-fold) is a widely-used procedure (eg [Cho et al., 2010]) but may lead to failure to identify true associations [Abraham et al., 2013]. Since that was not an aim of this project, and pre-selection was liberal, allowing all SNPs with p-values $< 1 \times 10^{-3}$, this is unlikely to have adversely affected the predictive ability of the GRS in this case.

An important consideration in the choice of cross-validation procedure is the bias-variance trade-off [Friedman et al., 2001]. Given the small sample size and apparently small effect sizes from the Q-Q plots, I expected that a model fit to $< 80\%$ of the data would have substantially reduced predictive power compared to one fit to the whole dataset (downward bias). Correspondingly, the computationally complex procedure and necessity for permutation testing to estimate the null distribution of $m(Y, y_T)$ precluded less biased XV-procedures such as leave-one-out cross validation. The ten-fold procedure was a reasonable compromise, although still computationally complex, requiring several thousand hours of computational time. The number of variants included in the GRS was chaotic across XV-folds (table 6.5), suggesting that the sampling variance in optimal λ_0 is high (equation 6.6).

Application

Even the best GRS for EO/PO prediction had modest sensitivity and specificity (table 6.4), and these values are likely to be overestimates of performance in a new dataset. The accuracy of predictors is likely to increase as more data becomes available, and leveraged scores may become more accurate than EO/PO agnostic GRS. Even the small degree of predictive accuracy observed in this study may be clinically useful, again if the GRS contains information independent of that attainable clinically. In practical terms, children with oligoarthritis and high GRS values could be monitored more closely for extension than those with low GRS values.

Although GRS are not widely used in this way, many clinical measurements which are only slightly correlated with a phenotype remain part of standard diagnostic procedures. For example, diagnostic recommendations for rheumatic heart disease [Reményi et al., 2012] include considerations of low-accuracy predictors such as ethnicity, geographical location and living conditions. Indeed, the standard clinical history and physical examination - an indispensable component of diagnosis - can be statistically considered as a collection of a large number of variables with low individual predictive power [Wipf et al., 1999]. In general, medical decisions are typically made on the basis of combined consideration of many predictors, and new predictors, however small their effects, can still be useful. GRS in particular have a cost-benefit advantage in that genotyping ideally only needs to be performed once in the lifetime of an individual [Abraham et al., 2014].

In order to be useful as a clinical predictor, the GRS has to explain phenotypic variance additional to that explained by existing predictors. Equivalently, it must remain correlated with the phenotype conditional on other predictor variables. In the case of the EO/PO phenotype, although there are some clinical associations, predictability is modest [Al-Matar

et al., 2002, Gardner-Medwin et al., 2006] suggesting that the GRS may indeed improve predictive accuracy. However, I was not able to assess this as further phenotypic information was unavailable.

Summary

This project gave a promising (though not definitive) indication that EO/PO is partially predictable from patient genotypes. While larger and more diverse datasets are needed to confirm this predictability and enable it to be clinically useful, this provides a useful starting point and general framework for investigation of genetic predictability of disease subtypes.

6.4 Investigation of heterogeneity in JIA subtypes

6.4.1 Motivation and general methods

As discussed in section 6.1, the classification of JIA into subtypes is largely based on clinical presentation. Consequently it is not necessarily the case that subtypes represent different pathological forms of the disease, or that there are systematic genetic differences between subgroups. Understanding pathological subtypes of a disease is useful in further investigation of the disease, and in guiding further scientific investigation. As discussed in chapter 5, the determination of an 'optimal' partitioning of a phenotype on the basis of genomic data is a difficult problem. However, it is possible to gauge how well a given subgrouping differentiates patients genetically in various ways. Chapter 5 details one such method which effectively assesses how well-separated two subtypes of a disease are on the basis of associations between control samples and combined case samples for the disease.

In this section, I examine a similar problem, in analysing how well-separated disease subgroups are on the basis of associations between either JIA and controls, T1D and controls, or RA and controls. The aim in this chapter is different from the aim in chapter 5 in two ways. Firstly, the question of whether subgroup differentiation is pleiotropic with JIA/control, RA/control, or T1D/control is of minor importance. Although leverage is used to help the analysis, the main aim is to assess whether there is genetic differentiation between subtypes at all. Secondly, an important question is *how* subgroups separate, in the sense of assessing which subgroups are most genetically similar, which cannot be easily assessed using the model in chapter 5.

I generally considered the null hypothesis H_0 that variation in all genetic variants was independent of subtype status. For SNPs, this is equivalent to equality of expected values

of numeric genotypes across all subgroups. It also implies that test statistics fitted using genotype data without knowledge of subtype status are themselves independent of subtype status. Throughout this section, H_0 will refer to this hypothesis unless specified.

The main methodological idea in this section is to construct several discriminants y_i of the form

$$y_i = \sum_{j \in \text{SNPs}} w_{jk} g_{ij} \tag{6.24}$$

where g_{ij} is the genotype of sample i at SNP j, and k is the XV fold which i is part of. The values w_{jk} is agnostic to the values y_i for $i \in$ fold k. This ensures that for samples i within each fold k, y_i is independent of the JIA subtype of sample i under H_0 I then compare the values of y_i between individuals with different subtypes of JIA to attempt to reject H_0. This also allows determination of how each discriminant 'separates' the subtypes by assessing the relative directions in which the mean values of y_i in each subgroup differ from the overall mean of y_i.

6.4.2 Modified GRS for RA and T1D

I derived GRS scores for RA and T1D which were agnostic to sample labelling for samples in the JIA dataset (ie subtype/case/control). This ensured that scores were unbiased for predicting subtype status and JIA/control status; under the null hypothesis that JIA/control status has no genetic associations, if $GRS(i)$ is the GRS value for sample i:

$$E\{GRS(i)|i \in \text{JIA cases}\} = E\{GRS(i)|i \in \text{controls}\} \tag{6.25}$$

and under H_0 (although heritability may be nonzero for JIA/control status), we have, for any subtype X

$$E\{GRS(i)|i \in X\} = E\{GRS(i)|i \in \text{JIA cases}\} \tag{6.26}$$

The second of these conditions is more important, since in this section I primarily aimed to assess whether GRS could differentiate subtypes of JIA. The first condition additionally enables assessments to be made of whether these GRS can predict subtype/control status. I used similar methods to those used to develop GRS for T1D and RA in sections 6.3.3 and 6.3.3. However, the final GRS developed in those sections made use of JIA/control sample labels, and hence did not satisfy equation 6.25, although they did satisfy equation 6.26.

In the fitting of the RA GRS in section 6.3.3, the only step involving JIA/control status was the determination of the number of variants to include in the GRS. For this application,

I replaced this step with the simple criterion of including all variants with $p_{RA} < 5 \times 10^{-8}$ (after pruning variants for LD with $r^2 < 0.2$).

The GRS scores for T1D in section 6.3.3 made use of case/control labels both in determining the number of variants to include and determining the effect directions for each variant. For the current application, I computed GRS scores for T1D by randomly splitting the data into ten parts, then fitting effect directions to each nine-tenths of the data and computed GRS scores for the remaining tenth using these effect directions, again using all variants with $p_{T1D} < 5 \times 10^{-8}$ after pruning for LD.

For both analyses, I computed two sets of GRSs, one including and one excluding the MHC region (chr. 6, 25.0-40.0 Mb, NCBI build 37). In both analyses, I used a stringent p-value threshold for inclusion (5×10^{-8}) since the levering GWAS were both large and well-powered, and an effect size corresponding to $p_{RA} = 5 \times 10^{-8}$ or $p_{T1D} = 5 \times 10^{-8}$ would correspond to a much smaller p_{JIA} value if the effect size were conserved in JIA ($p_{JIA} > 0.5$ if $p_{RA} = 5 \times 10^{-8}$, $p_{JIA} = 1.5 \times 10^{-2}$ if $p_{T1D} = 5 \times 10^{-8}$).

For each subtype, I assessed subtype/control differences in GRS scores using signed statistics from standard two-group t-tests. In order to simulate under a null hypothesis of homogeneity of genetic effects in JIA subtypes, I repeatedly randomly permuted subtype labels (but not JIA control labels) and recalculated summary statistics Z' for each random permutation. I used these sets of permuted 'null' Z' to assess significance of each observed Z score.

6.4.3 Unsupervised GRS construction

In order to capture 'inherent' variance within the JIA phenotype, I used a principal component analysis (PCA) on JIA case genotypes, weighted by association with JIA/control status. In typical genomic data, most of the variance in the first several standard principal components corresponds to identity by descent, or population substructure [Price et al., 2006]. Allelic differences between populations in the UK tend to be small but widespread [Leslie et al., 2015], so in smaller sets of variants in which other sources of variation (for instance, presence or absence of disease) lead to larger allelic differences, the contribution of population substructure to the variance of the first principal component should be reduced. By weighting the variance of each variant according to association with JIA/control, attention is concentrated to a small set of variants for which the principal source of genetic variance may be subtype status. Under H_0, the first principal component will not be associated with subtype status. If there is between-subtype genetic variance at the variants contributing to JIA risk, the first

principal component should have different values between subgroups, although population substructure and other confounders may still contribute to its variance.

In a standard PCA, the scores t for the first principal component on a genotype matrix $G = g_{ij}$ (sample i, variant j) are defined as:

$$t_i = \sum_j w_j g_{ij} \tag{6.27}$$

with weights $w = \{w_j\}$ satisfying

$$\begin{aligned} w &= \arg\max_{||w||=1} \sum_i t_i^2 \\ &= \arg\max_{||w||=1} w^T G^T G w \end{aligned} \tag{6.28}$$

In this application, I transformed each column g_j of g_{ij} according to z_j, a z-score for SNP j for the JIA/control comparison:

$$g'_j = \sqrt{\frac{z_j^{JIA/Control}}{\text{var}(g_j)}}(g_j - \bar{g_j}) \tag{6.29}$$

by comparison to the standard transform for PCA to equality of sample variances for each variable:

$$g'_j = \frac{g_j - \bar{g_j}}{\sqrt{\text{var}(g_j)}} \tag{6.30}$$

so that each g'_j had mean 0 and variance z_j instead of variance 1. By doing this, I expected that variance would correspond to genetic differences between JIA and control rather than genetic differences due to population substructure or identity by descent.

I initially pruned SNPs by LD to $r^2 < 0.2$, prioritising by p-value for JIA/control, and removed all SNPs with $|z_j| < \Phi^{-1}\left(\frac{1\times10^{-3}}{2}\right)$ (equivalent to $p_j > 1 \times 10^{-3}$). The remaining SNPs were used to generate the first principal component t'. I repeated the analysis with the MHC region removed (co-ordinates as in the previous section) to generate the first principal component t'_{NMHC}.

Since the construction of t' and t'_{NMHC} is agnostic to subtype labels, their distributions are the same in each subtype under H_0. I assessed deviance of each subtype by comparing values in that subtype with values not in that subtype. I additionally assessed whether t' and t'_{NMHC} had different means across subtypes in general using standard one-way ANOVA.

6.4.4 Supervised GRS construction

Finally, I assessed whether a supervised predictor - a linear discriminant - could differentiate JIA subtypes when trained to do so. While PCA heuristically finds the axis along which the data has the greatest variance, LDA finds the axis along which the ratio of between-group variance to within-group variance is greatest. This can lead to very different behaviour of the two predictors (figure 6.8)

Assume that N samples are partitioned into K classes $c_1, c_2..c_K$. Denote by g_i the set of genotypes of individual i, and assume that genotypes are standardised to have mean 0 across all samples. Denote $\hat{\mu_k} = \frac{1}{|c_k|}\sum_{i \in c_k} g_i$ as the within-class mean genotype for class k. The within-class covariance matrix Σ_w is estimated as

$$\hat{\Sigma_w} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i \in c_k}(g_i - \hat{\mu_k})(g_i - \hat{\mu_k})^T \tag{6.31}$$

and the between-class covariance matrix Σ_b estimated as

$$\hat{\Sigma_b} = \frac{1}{N}\sum_{k=1}^{K}|c_k|\hat{\mu_k}\hat{\mu_k}^T \tag{6.32}$$

The (first) canonical discriminant [Rao, 1948] (CD) is a score $\beta^T g_i$ that satisfies

$$\beta = \arg\max_\beta \beta^T \hat{\Sigma_b}\beta \tag{6.33}$$

subject to $\beta^T\hat{\Sigma_w}\beta = 1$. I scaled the genotype matrix g_{ij} in the same way as in section 6.4.3, so that the variance in cases of genotypes at variant j was proportional to $|z_j|$, an absolute z-score for the difference between cases and controls.

Since LDA is a supervised procedure, an assessment of whether an LDA fitted to the whole dataset could discriminate subgroups within the same dataset would be severely biased upwards (favouring discrimination), even under H_0. In order to avoid this bias, I used a ten-fold cross-validation procedure, fitting the CD to each nine-tenths of the model and using the resultant value of β to compute values of the discriminant for the final tenth. I assessed discrimination of each subtype in the same way as in section 6.4.3.

6.4.5 Genetic associations with subtype status

Initially, I conducted an a SNP based analysis to detect systematic differences between subtypes of JIA. I used the same datasets as described in the previous section. I performed

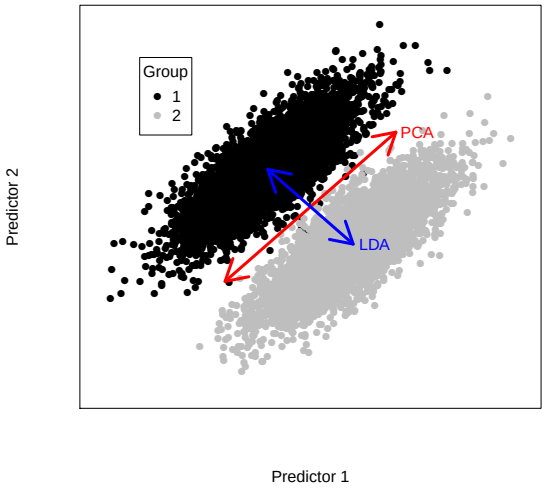


Fig. 6.8 Discriminants derived from PCA and LDA can behave markedly differently. Broadly, PCA corresponds to the axis of greatest variation, which would be expected to differentiate subtypes which differ in their fundamental disease mechanism. LDA finds the axis along which subtypes are best discriminated. Although the linear discriminant in this example roughly corresponds to the second principal component, and hence PCA alone may suffice, in higher dimensions the number of principal components which potentially have to be analysed to find one differentiating subtypes may be prohibitively high.

a GWAS using a one-way ANOVA for each SNP, comparing within-subtype variance to between-subtype variance, and computed p-values p_{diff} from the resulting F-scores. Seven subtypes were included in the analysis (those in table 6.2), considering EO and PO as separate, and excluding unclassified samples and samples with missing subtype labels.

Multiple SNPs reached genome-wide significance ($p_{diff} < 5 \times 10^{-8}$) and Bonferroni-corrected significance ($p_{diff} < 6.9 \times 10^{-9}$) in the MHC region. The lowest p-value was at rs146683910, at which $p_{diff} = 5.5 \times 10^{-72}$. The SNP rs146683910 was also associated with JIA/control status ($p_{JIA} = 1.3 \times 10^{-34}$) but not with EO/PO status ($p_{EO} = 0.22$). No SNPs outside the MHC region reached Bonferroni-corrected significance.

Substantial deviation from $p_{diff} \sim U(0,1)$ was evident in a Q-Q plot of $-\log_{10}(p_{diff})$ values with the MHC region included, but there was little evidence of inflation with the MHC region removed 6.9. The deviation from $U(0,1)$ increased when successively restricting to subsets of SNPs with stronger evidence of association with p_{JIA}, p_{RA}, and p_{T1D} (see chapter 2, section 2.2.6), more clearly when the MHC region was included. Conditional Q-Q plots are shown in figure 6.10.

I performed cFDR analyses on directly typed SNPs for subtype differentiation conditioned on T1D, RA and JIA. In all three analyses, a single SNP - rs2476601 - reached significance, using a significance threshold based on conserving FDR between a GWAS analysis and the cFDR procedure (threshold α_2, chapter 3, section 3.2.2. The SNP rs2476601 is on chromosome 1 near the *PTPN22* gene. It is possibly a causal variant in the region for RA [Stahl et al., 2010], T1D [Onengut-Gumuscu et al., 2014] and JIA [Hinks et al., 2013].

Collectively, these results indicate that there exist genetic variants differentiating JIA subgroups, namely in the MHC region and near *PTPN22*. There is also evidence of pleiotropy between subtype status and RA/control, T1D/control and JIA/control status, principally in the MHC region.

6.4.6 GRS results

Similarity between JIA subtypes and adult diseases

Results for the GRS fitted to RA are shown in table 6.6. Both GRS were significantly associated with JIA/control status ($p = 1.4 \times 10^{-9}$ with MHC included, $p = 1.8 \times 10^{-7}$ with MHC excluded). Under the null hypothesis that JIA subtypes are genetically homogeneous, the expected predictive accuracy of the GRS for determining subtype/control status should be equal for all subtypes, and the associated p-value (column 5 of table 6.6) should be monotonically decreasing with the number of samples in each subtype cohort. The subtype

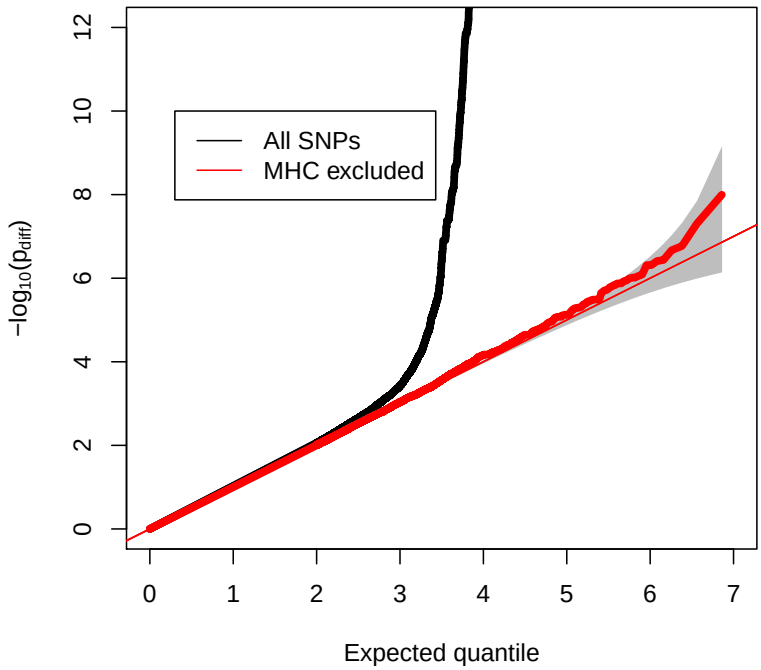


Fig. 6.9 Q-Q plot for p-values derived from F-scores for JIA subtype differentiation. A 99% pointwise confidence envelope is shown in grey, based on the distribution of order statistics from a uniform distribution. There is evidence of departure from a uniform p-value distribution when the MHC region is included, but not when it is removed.

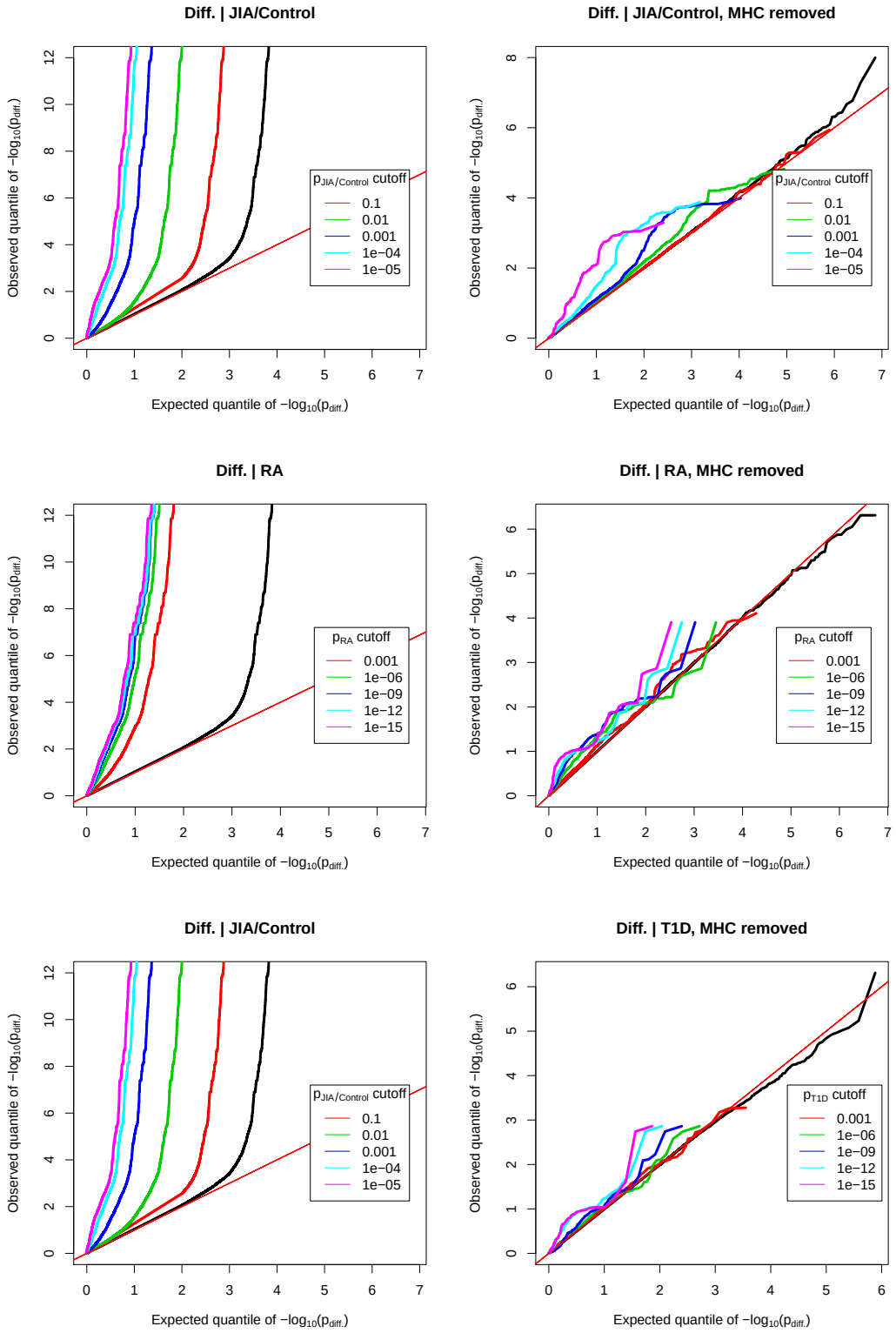


Fig. 6.10 Conditional Q-Q plots for inter-subtype differences in JIA, conditioned on JIA, RA and T1D, with MHC included (left panels) and excluded (right panels). Note different thresholds for conditional p-values, chosen because the studies on RA and T1D were larger than the study on JIA. There is evidence of deviation from a uniform distribution for p_{diff} when conditioning on association with other phenotypes, most clearly when the MHC regions is included. The reduced deviation when the MHC region is removed suggests that the inflation is largely MHC driven. Confidence envelopes are not shown as they differ for each conditional p-value threshold.

for which the GRS most strongly separated cases from controls was EO, for which prediction was significantly better in the observed data than under the null ($p < 1 \times 10^{-4}$; Bonferroni-adjusted p-value threshold 3.6×10^{-3} over 14 tests).

The GRS with MHC included could not distinguish ERA or PsA cases from controls (p-values for subtype/control comparisons 0.53, 0.76, 0.59 respectively), although the GRS with MHC excluded could predict RF+poly and PsA subtypes to an extent (p-values for subtype/control comparisons 3.28×10^{-3}, 3.10×10^{-3}). Since all variants in the second GRS were included in the first with the same odds-ratio, this suggests that the variance of the GRS with MHC included is dominated by the effects of SNPs in MHC which primarily affect heritability for EO, PO and RF-poly subtypes. Both GRS predicted Sys/control status poorly, but not much more poorly than would be expected if subtypes were genetically homogeneous.

The poor performance of both GRS on RF+poly was unexpected, as RF+poly JIA would seem to be the most phenotypically related to RA. The substantially improved predictability of EO/control status compared to other subtypes was also unexpected, since the subtype is phenotypically similar to the PO, RF-poly and RF+poly subgroups.

Results from the GRS fitted to T1D are shown in table 6.7. Both GRS were significantly associated with JIA/control status ($p = 5.0 \times 10^{-71}$ with MHC included, $p = 5.2 \times 10^{-10}$ with MHC removed). Again, the GRS with MHC included could predict EO/control status significantly better than what would be expected if subgroups were genetically homogeneous ($p < 1 \times 10^{4}$, Bonferroni-adjusted p-value threshold 3.6×10^{-3} over 14 tests). Prediction of systemic JIA was significantly worse than expected ($p < 1 \times 10^{4}$ with MHC included, $p = 3.8 \times 10^{-3}$ with MHC excluded) indicating that in a broad sense, the genetic determinants of Sys/control status are distinct from the determinants of T1D/control status. Given the genetic similarity of JIA and T1D [Onengut-Gumuscu et al., 2014] and the evidence of genetic differences between systemic and non-systemic JIA [Ombrello et al., 2017], this suggests that similarity between JIA and T1D arises primarily from associations with non-systemic JIA.

Unsupervised GRS

PC1 scores from the unsupervised GRS with MHC region included were significantly different in JIA subtypes (ANOVA, $p = 3.1 \times 10^{-12}$). This is consistent with the results from single-SNP analyses in section 6.4.5, again indicating evidence for genetic differences between JIA subtypes.

A plot of densities of PC1 scores by subtype is shown in figure 6.11. This visually suggests that the between-subtype discrimination is driven to a large extent by the Sys

Table 6.6 Details of GRS fitted to RA for differentiating JIA subtypes. Column 2 is the number of samples in each subtype. Column 3 is the difference in mean GRS scores (normalised to overall mean 0, variance 1) between subtypes and controls. Column 4 is a p-value derived from a t-test against the null hypothesis that the GRS is independent of case/control status. Column 5 is the probability that an observed t-score would be further than the observed score from the mean of t-scores derived from random relabelling of subtypes, and can be interpreted as a p-value against the null hypothesis implies that the accuracy of prediction of that subtype from controls by GRS is equal to the accuracy of prediction of JIA samples from controls on average. Column 6 indicates whether the differentiation of that subtype from controls was worse or better than the average differentiation of JIA samples from controls.

Subtype	**Cases**	**Mean GRS dif.**	**P (sub/ctl)**	**P (vs pred)**	**Dir.**
MHC included					
Sys	280	0.14	0.023	0.77	Worse
PO	650	0.19	1.17×10^{-05}	0.37	Better
EO	390	0.4	5.44×10^{-14}	$< 1 \times 10^{-4}$	Better
RF-poly	570	0.12	7.52×10^{-03}	0.32	Worse
RF+poly	200	-0.044	0.53	3.80×10^{-03}	Worse
ERA	180	-0.023	0.76	0.01	Worse
PsA	150	0.054	0.59	0.15	Worse
MHC excluded					
Sys	280	0.062	0.3	0.2	Worse
PO	650	0.12	9.51×10^{-03}	0.48	Worse
EO	390	0.12	0.023	0.77	Worse
RF-poly	570	0.21	3.28×10^{-05}	0.13	Better
RF+poly	200	0.24	1.67×10^{-03}	0.13	Better
ERA	180	6.42×10^{-03}	0.93	0.066	Worse
PsA	150	0.24	3.10×10^{-03}	0.12	Better

Table 6.7 Details of GRS fitted to T1D for differentiating JIA subtypes. Column 2 is the number of samples in each subtype. Column 3 is the difference in mean GRS scores (normalised to overall mean 0, variance 1) between subtypes and controls. Column 4 is a p-value derived from a t-test against the null hypothesis that the GRS is independent of case/control status. Column 5 is the probability that an observed t-score would be further than the observed score from the mean of t-scores derived from random relabelling of subtypes, and can be interpreted as a p-value against the null hypothesis implies that the accuracy of prediction of that subtype from controls by GRS is equal to the accuracy of prediction of JIA samples from controls on average. Column 6 indicates whether the differentiation of that subtype from controls was worse or better than the average differentiation of JIA samples from controls.

Subtype	**Cases**	**Mean GRS dif.**	**P (sub/ctl)**	**P (vs pred)**	**Dir.**
MHC included					
Sys	280	0.21	1.08×10^{-03}	$< 1 \times 10^{-4}$	Worse
PO	650	0.45	2.67×10^{-25}	0.6	Worse
EO	390	0.68	1.61×10^{-36}	$< 1 \times 10^{-4}$	Better
RF-poly	570	0.45	3.41×10^{-25}	0.84	Better
RF+poly	200	0.44	7.87×10^{-10}	0.88	Worse
ERA	180	0.51	1.23×10^{-10}	0.67	Better
PsA	150	0.38	2.33×10^{-06}	0.42	Worse
MHC excluded					
Sys	280	-6.5×10^{-4}	0.99	3.80×10^{-03}	Worse
PO	650	0.20	1.04×10^{-05}	0.39	Better
EO	390	0.21	1.36×10^{-04}	0.32	Better
RF-poly	570	0.22	3.10×10^{-06}	0.15	Better
RF+poly	200	0.28	1.97×10^{-04}	0.083	Better
ERA	180	0.036	0.64	6.80×10^{-03}	Worse
PsA	150	0.16	0.05	0.93	Better

subtype. Tests of discrimination between each subtype and the remainder of the cases are shown in table 6.8, which confirm this. PC1 scores for the Sys subtype were significantly different from the JIA average ($p = 4.68 \times 10^{-10}$, not adjusted for multiple testing). The EO subtype was again clearly discriminated from the average ($p = 4.91 \times 10^{-8}$, not adjusted), with deviation in the opposite direction from the JIA mean to PC1 scores for Sys JIA. The distribution of PC1 values for the ERA subtype appeared to be shifted from the mean scores in the same direction as the EO subtype (figure 6.11) but evidence for this was inconclusive ($p = 0.089$).

If we consider individual genotypes as points in a high-dimensional space in which each dimension corresponds to a SNP, then the procedure of weighting variances according to equation 6.29 is equivalent to a linear transformation on this space where the associated matrix is diagonal, and the transformation is chosen to maximise the separation between JIA samples and controls in the new space. The results above indicate that along the axis in which points corresponding to JIA samples in this transformed space have the greatest variance (PCA), EO and Sys samples tend to fall on opposite sides of the mean from the JIA average.

PC1 scores with the MHC region removed were not able to differentiate subtypes in general ($p = 0.70$) and no individual subtype was significantly differentiable from the mean (table 6.7). This was relatively unsurprising, given that no individual non-MHC variant reached Bonferroni-corrected significance for differentiating subtypes, and there was no visible inflation of p-values on a Q-Q plot when the MHC region was removed (figure 6.9). Although there was some evidence of inflation leveraged on JIA (figure 6.10) this was not reflected in the GRS.

Supervised analysis

The cross-validated CD scores for which MHC was included had different means across subtypes (ANOVA, $p < 3 \times 10^{-37}$). This was consistent with findings from the PC-based GRS and the single-SNP analyses. The CD-based score was better able to differentiate subtypes than the PC-based score, possibly due to being fitted specifically to differentiate subtypes rather than to characterise the maximum variance in JIA genetics. Plots of densities of CD scores are shown in figure 6.12 and results are tabulated in table 6.9. The CD with MHC included showed striking differentiation between ERA samples and the JIA average ($p = 3.3 \times 10^{-20}$), and again showed differentiation of the EO subtype from the JIA average. The CD scores with MHC excluded did not have significantly different means across subtypes

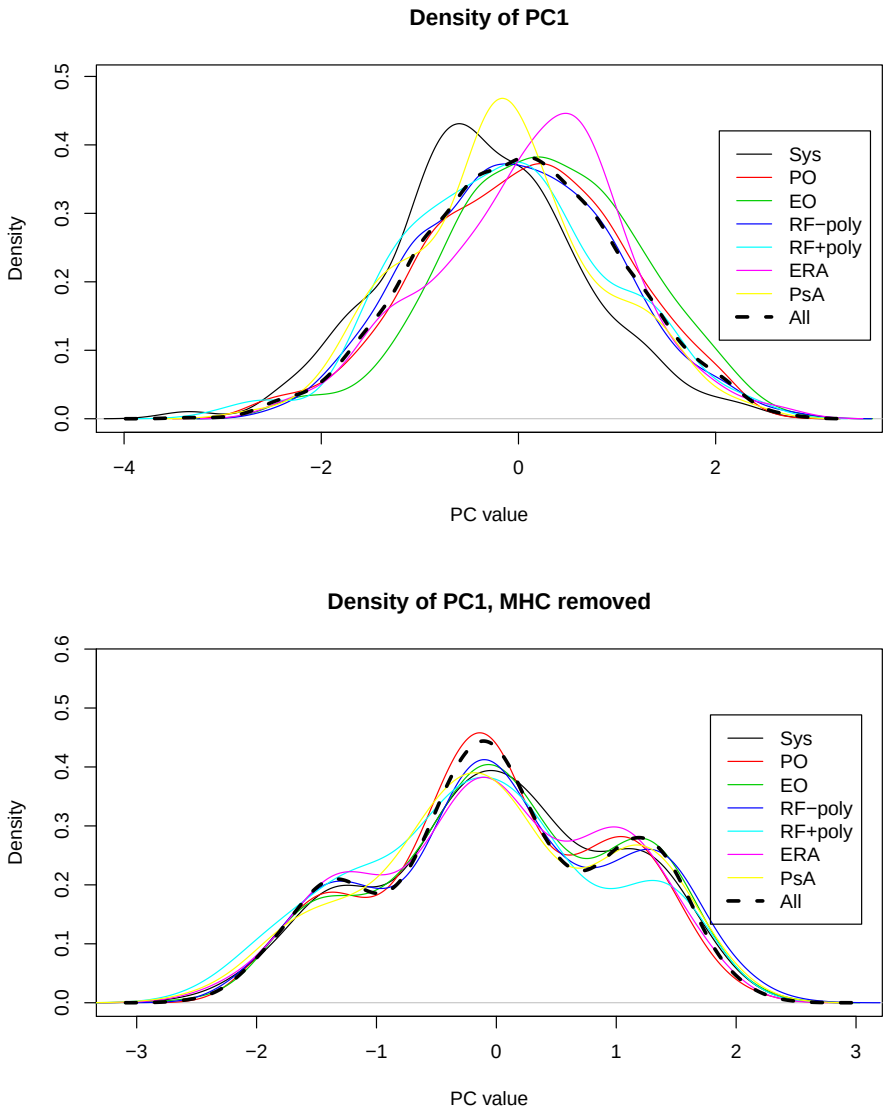


Fig. 6.11 Densities of PC1 for JIA subtypes, where PCs are computed weighting on JIA/control status. Some evidence of separation of subgroups is seen, particularly for systemic JIA (black solid line) and EO (green). The apparent trimodal distribution in the lower panel is due to the small number of variants in the risk score.

Table 6.8 Discrimination of JIA subtypes using the first PC weighted by z-score for JIA/control. Column 'Mean PC1 dif' is the difference in average PC1 score for each subtype from the overall mean PC1 score for JIA samples (note that these differences are not independent). PC1 scores for Sys and EO subtypes are strongly differentiated from the JIA overall average in the analysis with MHC included. Notably, PC1 scores for Sys and EO subtypes deviate from the mean PC1 score in opposite directions.

Subtype	**Cases**	**Mean PC1 dif**	**P value**
MHC included			
Sys	280	-0.39	4.68×10^{-10}
PO	650	0.061	0.18
EO	390	0.29	4.91×10^{-08}
RF-poly	570	-0.022	0.64
RF+poly	200	-0.12	0.1
ERA	180	0.13	0.089
PsA	150	-0.15	0.065
MHC excluded			
Sys	280	0.017	0.79
PO	650	-0.019	0.66
EO	390	0.058	0.29
RF-poly	570	0.052	0.28
RF+poly	200	-0.13	0.088
ERA	180	-0.028	0.71
PsA	150	-9.59×10^{-03}	0.91

(ANOVA, $p = 0.88$) and no subtype had a significantly different mean CD score compared to the other subtypes.

In the transformed space described in the previous subsection, the canonical discriminant corresponds to the axis along which the predefined subgroups are best separated, as opposed to the axis along which variance across samples is greatest (see figure 6.8). This suggests that along this axis, ERA is strongly separated from other subtypes. It is not as well-separated on the axis corresponding to the first PC.

6.4.7 Discussion

This series of analyses demonstrated evidence that genetic risk scores can be constructed which differentiate certain JIA subgroups, principally involving variants in the MHC region, and that JIA subtypes are differentiated in different ways according to the method used to weight variants.

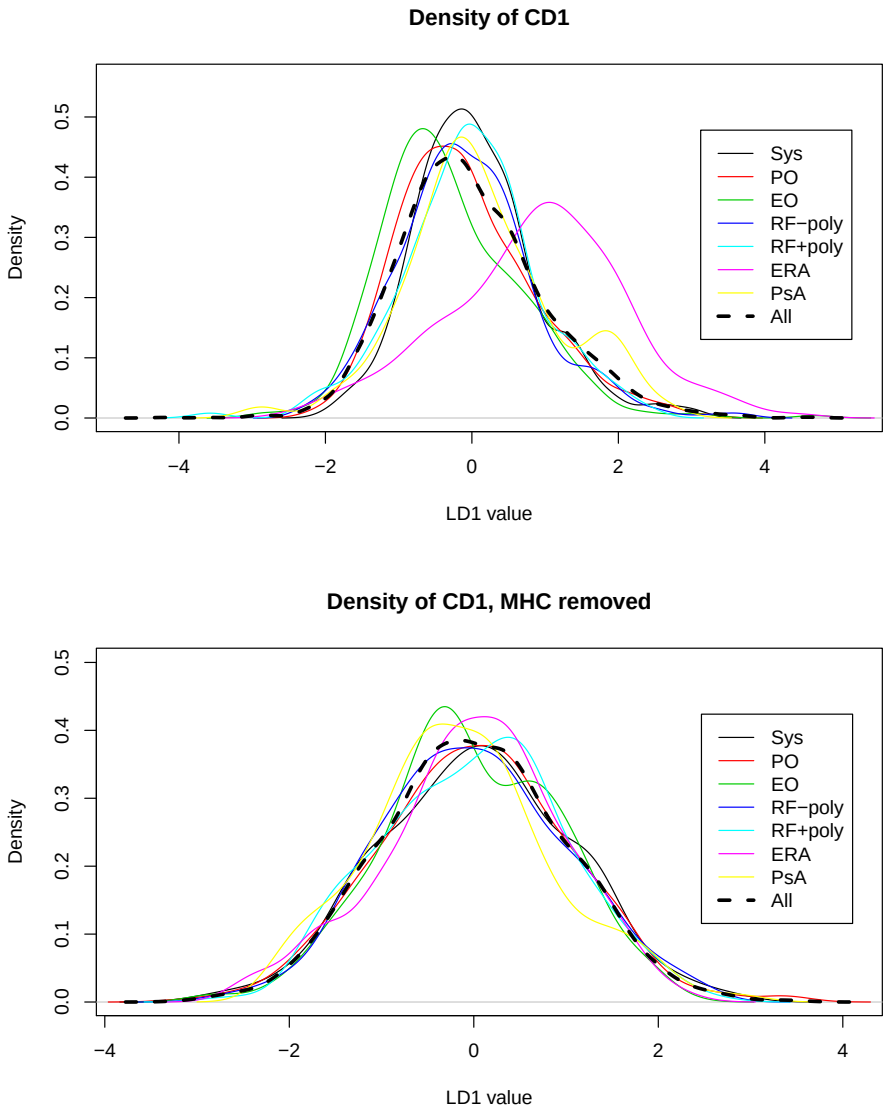


Fig. 6.12 Densities of unbiased canonical discriminant for JIA subtypes, where CDs are computed weighting on JIA/control status. Some evidence of separation of subgroups is seen, particularly for the ERA (purple) and EO (green) subtypes.

Table 6.9 JIA subtypes using the first CD weighted by z-score for JIA/control. Column 'Mean CD1 dif' is the difference in average CD1 score for each subtype from the overall mean CD1 score for JIA samples (note that these differences are not independent). Mean CD differences are correlated, but this effect is small. CD1 scores for EO and ERA subtypes are strongly differentiated from the average scores for JIA in the analysis with MHC included. The deviation of CD1 is in opposite directions for EO and ERA.

Subtype	**Cases**	**Mean CD1 dif**	**P value**
MHC included			
Sys	280	0.06	0.27
PO	650	-0.098	0.023
EO	390	-0.34	1.37×10^{-10}
RF-poly	570	-0.092	0.041
RF+poly	200	-0.023	0.74
ERA	180	0.94	3.30×10^{-20}
PsA	150	0.12	0.16
MHC excluded			
Sys	280	0.029	0.65
PO	650	0.027	0.56
EO	390	-0.034	0.51
RFneg	570	1.05×10^{-03}	0.98
RFpos	200	6.04×10^{-03}	0.93
ERA	180	-0.011	0.88
PsA	150	-0.14	0.091

The improved predictive ability of a GRS for RA in the EO subtype suggests causative processes may be shared to a greater extent between EO and RA than between other JIA subtypes. This is consistent with the clinical observation that a symmetric distribution of affected joints is associated with oligoarthritic extension [Al-Matar et al., 2002]. The GRS for RA with MHC included was significantly less able to predict RF+poly than would be expected if subtypes were genetically homogeneous, but with MHC excluded, a GRS for RA was able to differentiate EO from controls about as well as expected under a null hypothesis of genetic homogeneity. This suggests distinct disease mechanisms for EO and RA associated with the MHC region, but potentially shared mechanisms controlled by elsewhere in the genome.

The GRS for T1D also predicted EO better than expected. Prediction of systemic JIA was significantly worse than expected, and the poor prediction remained in the GRS with the MHC region removed. This suggests an absence of shared mechanisms between T1D and systemic

JIA, which is notable given the good prediction of other JIA subtypes. The differentiation between systemic JIA and other subtypes was also observed in the unsupervised analysis using PCA, suggesting a fundamental difference in the mechanism of systemic JIA compared to other subtypes. This is consistent with earlier findings [Ombrello et al., 2017].

The supervised analysis (LDA) was notable for the sharp discrimination of ERA. The ERA subtype appeared to be moderately differentiated from other JIA subtypes on the basis of the GRS for T1D and the PCA analysis, but to a much lesser extent. This suggests a distinction between causative mechanisms of ERA and other JIA types, and highlights the importance of considering both types of discriminants. Further analysis of ERA could involve leverage on summary statistics for ankylosing spondylitis, a phenotypically similar adult disease [Colbert, 2010].

As discussed in the introduction to this chapter, an important question regarding the ILAR subtypes is whether they are the most clinically-relevant way to sub-classify the disease. The results from the single-SNP analysis and the unsupervised GRS (PC-based) lend credence to the current subtyping criteria by demonstrating the presence of genetic heterogeneity between ILAR subtypes, in addition to that demonstrated for Sys JIA [Ombrello et al., 2017]. In particular, the variation across subtypes observed in the first re-weighted principal component indicates that the current subtyping may capture the 'fundamental' variance in JIA genetic architecture to an extent. However, I noted that the between-subtype discrimination using the canonical discriminant was larger than that using PCA. If the ILAR subtyping corresponded to a maximum separation of subtypes in the transformed space described in section 6.4.6, then (under the assumption that the variance in PC1 is due to inter-phenotypic heterogeneity rather than identity by descent or other confounders) we would expect the PCA and LDA axes to be approximately the same, and the separation of subgroups along these axes to be similar. This suggests that other sub-classifications may better differentiate JIA cases on a genetic basis than the current ILAR subtypes.

The risk scores developed in this section do not have the immediate potential clinical application of those developed to differentiate EO and PO. Indeed, the findings of this section may appear vague, and it is not clear why an assertion of 'increased chance of similar disease mechanisms' should be useful either clinically or scientifically. Since studies of this sort do not definitively associate any given pathway with any JIA subtype of interest, further studies will be necessary in order to translate any assertion of disease similarity into a clinically useful knowledge of disease pathology.

However, there are a vast range of processes potentially contributing to arthritis and articular disease. In order to understand the action of a pathological process on a disease

in broad terms, the activity of that process has to be monitored in individuals with the disease. This is an expensive process, both in investigative resources and burden on the patient population. The identification of probable shared processes between diseases enables improved prior knowledge of the processes likely to be involved. By establishing similarities with well-studied and common disease such as RA and T1D, knowledge already attained about the common diseases can be used to reduce the burden of investigation in the rare one. The PC- and CD- based analyses contribute to understanding the way subtypes separate on the basis of JIA association itself, and may be useful in informing further clinically-useful disease subtypings. This can ultimately contribute to improving treatment provision in patients with JIA.

Chapter 7

Discussion

7.1 Conclusions and linking themes

In this thesis, I developed a range of methods for the comparison of case-control studies, with application principally in the analysis of genomic data for complex diseases. This entailed an important extension of a method for levering GWAS summary statistics in an asymmetric way (chapter 2), with application to the analysis of underpowered datasets (chapter 3); the development of a similar shared-control design in the context of study replication (chapter 4); the development of a characterisation of patterns of genetic heterogeneity in disease subtypes (chapter 5) and the specific investigation of heterogeneity patterns in JIA using predictive risk scores (chapter 6). Recalling figure 1.1 in chapter 1, these topics are not sequential and their goals are varied; however, they are linked by several underlying ideas.

Individual methods and findings are discussed in the relevant chapters. In this chapter, I will begin by discussing the linking themes of this work, and their advantages and shortcomings. I will then examine the wider context of this work and discuss how it can contribute to scientific understanding of the genetics of complex disease. Finally, I will discuss the implications of this work in the development of precision and personalised medicine, and suggest several areas for further research.

7.1.1 Joint analysis of two traits

Every project in this thesis concerns the co-analysis of more than one trait or study; chapters 2 and 3 concern analysis of explicitly different diseases, chapter 4 different studies on the same disease, and chapters 5 and 6 studies relating to heterogeneity within a disease.

The advantage of co-analysis of related traits, as opposed to single-trait analyses or meta-analysis of both traits together, is improved power for detecting association in one study while minimally compromising the specificity of the trait under investigation. If one trait is analysed at a time, this improved power is lost; if traits are meta-analysed as though they were the same disease, it is difficult to ascribe a resulting association specifically to one of the two traits.

However, the compromise is imperfect; some specificity and power are necessarily lost. As discussed in chapter 3, section 3.4, the ordering of variants by $\widehat{cFDR}$ is systematically different to the ordering of variants by p-value when there is pleiotropy between the principal and conditional phenotypes. Disregarding LD, the expected p-values of variants of a fixed AF are monotonic to the effect sizes of those variants, but this is not true for the expected $\widehat{cFDR}$ values. The same is generally true for any procedure assessing association with one phenotype conditioning on another. For instance, a test for association could be developed based on the fitted coefficient for a SNP of interest in a GRS for the phenotype of interest levered on a second phenotype such as those in chapter 6, section 6.3.3, but such a test would be influenced both by association with the phenotype of interest and by association with the levering phenotype. Thus, although the cFDR procedure and other association tests using leverage can be used to identify the set of regions associated with a disease, it cannot be used directly to assessing the relative effect sizes of variants.

If large enough datasets were available, co-analysis of traits would not be necessary, and the compromises could be avoided. 'Large enough' is nonetheless intractable; even as sample sizes increase through the hundreds of thousands of samples, new associations with complex traits can still be discovered [Wood et al., 2014]. It is impossible that anywhere near this power may be attainable when analysing rare diseases, and even if it were, the trend of subdividing diseases into ever-finer subtypes [Robinson et al., 2008] will probably continue as interest in 'precision medicine' develops [Collins and Varmus, 2015], so interest in small studies is likely to persist. Furthermore, variant effect sizes for the same disease may differ between cohorts of patients with different ethnicities, and as GWAS expand in scope past the western world, comparison of results between populations will continue to be of medical interest [Rosenberg et al., 2010].

The promise of co-analysis of two studies suggests a natural extension to more than two studies. This extension could lead to deeper understanding of genetic relationships between diseases; for example, the similarities and differences between JIA prediction using T1D and RA in chapter 6, and the pleiotropy explored in chapter 3 suggests the presence of categories of variants affecting only two of the three traits as well as all three. However, three- or

more- trait analysis becomes difficult in terms of data requirements. All the methods for joint analysis in this thesis require modelling distributions of effect sizes in some way, but this becomes difficult for more than two diseases. In practical terms, given three traits i, j, k, a cFDR analogue of the form

$$Pr(H_0^i | P_i < p_i, P_j < p_j, P_k < p_k) \tag{7.1}$$

would require estimation of densities of the three-dimensional random variable (p_i, p_j, p_k), which would become very erratic in regions where observations are sparse. However, this may be an interesting avenue of further research, if sufficiently detailed datasets are available.

7.1.2 Adaptations to a shared control design

A second common theme in this thesis is the comparison of studies with shared controls. As derived in chapters 2 and 4, sharing controls (or cases) induces a correlation between resultant effect sizes. This is managed in different ways in different chapters; in 2, p-values are transformed to account for the correlation; in 4, thresholds are adjusted to maintain false-positive rates, and in 5 (in which all controls are 'shared' between analyses for each subgroup), alternative summary statistics are used which are independent under the relevant null hypotheses.

The obvious advantage of sharing controls is enlargement of the control cohort, increasing power in univariate analyses (standard GWAS) and usually increasing power in bivariate analyses. Chapters 2 (figure 2.2) and 4 (figure 4.2) explore the improvement in power in several specific contexts. A disadvantage of sharing controls is the requirement for more complex analyses, or the use of alternative summary statistics which are harder to interpret.

Chapter 4 explores more subtle advantages and disadvantages in the context of 'aberrance' in various cohorts. If some error in sampling or measurement leads to a systematic difference between the expected value of some variable in a control cohort and the expected value of the variable in the population the cohort ostensibly samples, then sharing the control cohort between both studies will propagate this error into both studies. This could lead to loss of control of type-1 error rate in the study replication setting (chapter 4) or inappropriate leverage in the cFDR setting. In the cFDR setting in particular, errors in shared controls could lead to loss of control of type 1 error rate in two ways; firstly, by lowering p-values at affected variables in both the principal and conditional phenotypes, and secondly by suggesting pleiotropy where there may not be any, inappropriately lowering the p-value threshold for association at low conditional p-value thresholds, and increasing type 1 error

rate at unaffected variables. Conversely, sharing controls can also improve control over type-1 error rate if there is aberrance in other cohorts, explored in detail in chapter 4. This is of importance if there is differential confidence in the representativeness of cohorts of their respective populations.

Sharing controls is becoming common as datasets become more widely shared (eg [The Wellcome Trust Case Control Consortium, 2007, UK Biobank, 2007, Canela-Xandri et al., 2017]). Large GWAS meta-analyses frequently use controls from a range of sources, not necessarily recruited specifically for the study (eg [Okada et al., 2014, Barrett et al., 2009, Trynka et al., 2012], meaning that control subjects in such studies frequently overlap. This can be a problem if comparisons between studies are performed naively, since p-values are dependent between studies even for null SNPs. This provides a second important incentive for the development of shared-control designs in association study comparison: the enabling of type-1 error rate control in comparisons between large published datasets without the need to recalculate summary statistics.

7.1.3 Using multi-SNP effects

Chapters 5 and 6 both involve statistical inference based on multiple variants, rather than single-SNP effects. Chapters 2 and 3 search for single-SNP effects, but use the joint distribution of effect sizes of genome-wide SNPs to inform association tests.

All methods assume that the genetic architecture of the disease in question involves a large number of SNP-disease associations, with an approximately normal distribution of effect sizes (appendix A.1.1 to chapter 2 includes an exploration of robustness to deviation of effect sizes from a normal distribution). The method in chapter 5 is based on the assumption that there is greater power to detect multi-SNP effects than to detect than single-SNP effects; see figure 5.4.

From a biomedical perspective, the assumption of a large number of associated variants is justifiable. Biochemical and metabolic pathways are complex, and many perturbations could potentially lead to exacerbation of a pathological process. Variants strongly associated with a disease are expected to be under negative selection, so disease-associated variants typically have small effect sizes or are rare [Gibson, 2012]. An implication of this is that individual SNP-disease associations are generally difficult to find, so analysis of combinations of variants may be necessary in order to understand disease genetics in any way.

A problem with multi-SNP effects is difficulty in interpretation, with the result that the best follow-up for investigating a positive result may be unclear. Positive results using the

PLR test developed in chapter 5, for example, do not immediately suggest any lines of follow-up (other than the single-SNP tests outlined in section 5.2.5, although they do not require a positive result from the PLR test in order to be used). The clinical or scientific usefulness of 'similarity' of subtypes of JIA based on various discriminants in chapter 6 is also not immediately obvious. The methods do, however, have important scientific uses, which are discussed specifically in sections 6.4.7 and 7.1.5.

In general, many standard analyses of genomic data are focussed on finding single-SNP effects because the long-term goal is to identify new drug targets (eg [Wolfs et al., 2009, Zhang et al., 2008]). However, this is not the only way that genomics can contribute to medical practice. An important prospect for clinical contribution is the use of genetic risk scores as predictors of disease risk or disease outcome, as discussed in chapter 6 (though the extent is debatable [Clayton, 2009]). I propose that determination of the likely cause of subgroup heterogeneity in a disease (chapter 5) may also be useful in directing further research on heterogeneity to investigating predominantly environmental, physiological or pathological causes.

7.1.4 Efficient use of information

It is generally an important task in statistics to determine the best procedure to use for a given problem. A typical metric used for performance in frequentist statistics is the maximisation of power for a fixed type-1 error rate, as used in (for example) the Neyman-Pearson lemma. Determination of the exact or asymptotic power of a procedure in general circumstances is often mathematically intractable, particularly when trying to predict performances on complex structures such as genomic data. However, a useful heuristic for performance is whether the procedure incorporates everything we (the researcher) know about the data.

This heuristic is useful in several chapters. Chapter 2 is an adaptation of an existing procedure to make use of the information that associated variants may be common to related diseases (that is, show pleiotropy), which is otherwise not used in a standard GWAS analysis. I extended it to allow for shared control samples, allowing for incorporation of the knowledge that the expected value of variables under investigation is (generally) the same in both control sets. This is also a key part of the incentive for the procedure in chapter 4.

A more subtle argument in the cFDR method concerns its comparison against a more typical leverage procedure, in which only variants reaching a given p-value threshold in some conditional phenotype are tested in the principal phenotype, with a threshold on p-value for the principal phenotype determined using a Bonferroni correction (eg [Plagnol et al.,

2011]). This has the advantage over the cFDR of easy control over the overall FDR (using the Benjamini-Hochberg procedure) and an obvious advantage in simplicity of method. However, it means that variants narrowly missing the conditional p-value threshold are excluded from investigation, and variants with p-value thresholds reaching the threshold are treated the same way, no matter how associated they are with the principal phenotype.

In this case, the information from the summary statistics for the principal phenotype is reduced from a real-valued variable to a binary variable, representing a large sacrifice of information. The reduction in information is slightly smaller if the threshold is chosen according to the distribution of p-values for the conditional phenotype, but choosing a threshold in this way is difficult without biasing results (the 'Texas Sharpshooter problem'). The cFDR requires no such sacrifice, and should in general circumstances have greater power to detect variants at a fixed false-positive rate than methods based on a single-SNP cutoff.

The methods in chapters 5 and 6 both involve assessing within-subgroup association. The single-SNP analysis using the cFDR in chapter 5 and the PC/LD scores in chapter 6 both perform this analysis by levering on associations between the general disease and the control group (Z_a, using the notation of chapter 5). In this way, more information is used than would be under an isolated analysis of subtype differentiation.

An important caveat of increased information use is the question of whether the increased information actually helps. In an analysis of T1D|PSO in chapter 2, no inflation is evident for T1D after conditioning on association with PSO. In this case, intuitively, the inclusion of PSO data in the analysis of T1D adds no extra information to the analysis, and should not improve power to detect association with T1D at all. The analogue of this in the inter-subgroup comparisons of chapters 5 and 6 is independence between subgroup associations and case/control associations - the null hypothesis in chapter 5. This is one reason why assessment of this null hypothesis is important

The heuristic of information use is not perfect. As demonstrated in chapter 4, the improvements in power from using more information may be very small, and may come at the cost of losing control of other error rates. However, I consider it an important consideration in experimental and analytical design.

7.1.5 Genetic analysis of subgroups

In the introduction to this thesis, I proposed that the fields of medicine and medical research have changed markedly in the past several centuries in the way that they classify diseases. In the earliest medical records, patients are classified very specifically, only occasionally

drawing links between similar presentations. As medical science and society progressed, classifications of diseases grew broader, allowing greater power in assessing pathology and treatment strategies. Finally, in the past several decades, the trend has been reversed, and diseases are becoming ever more finely sub-classified. The merits of analysing diseases as large but heterogeneous cohorts or small homogeneous cohorts is known as the 'lumping-splitting' debate [McKusick, 1969].

The general reason for the subdivision of diseases is the hope for more specific treatment, avoiding the use of unnecessary therapy. From a clinical perspective, patients are often managed by successively trialling different therapies until one is effective (eg [Ravelli and Martini, 2007]). The advantages of predicting response to therapy are thus in reducing the time patients need to spend trialling therapies, and reducing the cost to the medical system from temporary trials of ineffective medication. The advantages are naturally greatest for medications which are effective in only a small (but identifiable) subset of disease cases, and are expensive to produce.

From a practical perspective, sub-classifications of disease are used to characterise patients who respond to a given therapy in a clinical trial. However, this is not simple. For a cohort of N patients, there are 2^N ways to select a patient subgroup. Even in a realistic clinical context, there are a huge number of potentially clinically useful subgroupings; for instance, by the presence or absence of any disease symptom or set of symptoms, by ethnicity or country, by body habitus or physiological parameters, or by aspects of disease course. The vast scope of potential subgroupings means that prior information on which subgroupings are likely to be able to predict treatment response must be incorporated. This can be done either by only considering subgroups likely to be predictive of treatment response (that is, using a prior which has probability 0 for unused subgroups), or explicitly constructing a prior over potential subgroupings.

The methods developed in chapters 5 and 6 are useful in producing such a prior. In particular, the methods in chapter 5 are important in differentiating the types of heterogeneity present in a subgrouping. This can enable more effective classification of patients who respond to a given therapy, ultimately improving the efficiency of the medical system and reducing harm to patients.

7.2 Future directions

7.2.1 Conditional analysis

An important application of several methods in this thesis is the use of a large dataset to enable a better analysis of a small one. In both chapter 2 and 6, the analysis is asymmetric, in that the hypotheses of interest only concern the small dataset; chapter 3 the cFDR methodology is used to investigate EGPA conditioning on a much larger dataset of GWAS association statistics for eosinophil count in healthy individuals, and in chapter 6, association statistics from a comparatively large study (RA/T1D/JIA) are used for both variable selection and coefficient estimation in GRS for a smaller dataset on a rare phenotype (EO/PO). In both the cFDR methodology and the methods in chapter 6, the association of variants with the levering phenotype does not need to be investigated, and indeed, the phenotype used for leverage does not need to be meaningful in itself.

This invites the possibility of assembling large GWAS datasets for the sole purpose of leverage. An example may be an assimilation of all cohorts of autoimmune disease cases in chapter 2. Given the widespread pleiotropy between autoimmune diseases, some of which was explored in the chapter, this could enable the identification of variants which have too small an effect size in any single phenotype. Importantly, a genetic association in such a case cohort would mean little on its own, since the evidence of association with any one disease may be poor - but the dataset as a whole could be an effective way to prioritise autoimmune-associated regions for analysis. The ImmunoChip and similar custom arrays were designed using similar reasoning.

The assembly of multiple case cohorts into one risks reducing effect sizes for variants only associated with a subset of the constituent cohorts. This reduction of effect sizes may be overcome by using a p-value based meta-analysis, as variants may have effects in opposite directions in different cohorts ([Cotsapas and Hafler, 2013]). Another problem may be that a meta-analysis of many diverse case-cohorts could identify too many associated variants to effectively reduce dimensionality. Despite these potential shortcomings, I believe that some assembly of some such cohorts could strengthen the power of standard GWAS analysis and facilitate genomic analysis of rare disease in cases where it may not previously have been possible. This is an example in which both the ‘lumping’ and ‘splitting’ philosophies could be used together.

7.2.2 Further characterisation of heterogeneity

The methodology presented in chapter 5 aimed to differentiate two particular classes of heterogeneity between disease subtypes; namely, testing for the presence of a set of variants associated both with subtype status and case/control status. As detailed in appendix D.1, table D.1, this dichotomises a larger set of potential genetic architectures of disease subtypes. A potentially useful line of further research would be to develop methods to differentiate two-dimensional genetic architectures further.

There are several metrics which could be used to develop such a characterisation process. An important starting point is assessing deviation of Z_d from an $N(0,1)$ distribution, to identify whether disease heterogeneity is completely environmentally-driven, with no heritability. This is more difficult that it may initially seem. The EO/PO phenotype in chapter 6 is an example; there is no evidence of deviation of p_{EO} from a uniform distribution, but evidence of deviation can be seen conditioning on other phenotypes, indicating that the EO/PO phenotype has a degree of heritability. Other important metrics, explored in appendix D.1, table D.1, include genetic correlation and the proposed 'absolute correlation'.

Another important potential application of the subgroups methodology is in the investigation of differential drug response. Different responses to therapy may be due (in varying degrees) to individual differences in pharmacokinetics and to different disease mechanisms, and the question of which of these is occurring is of importance in personalised medicine. As discussed in section 7.1.5, if the goal is to characterise a subgroup of patients likely to respond to a drug, a strong prior on likely subgroupings of patients is necessary. The methodology of chapter 5 applied to this problem could determine if the best subgroupings to look at are on the basis of disease symptomatology or individual pharmacokinetic and physiological variation.

7.2.3 Personalised, precision, and 'ballpark' medicine

A large part of the motivation for the topics in this thesis, and the field of genomics in general, is the development of precision medicine. This can be seen as the eventual goal of the movement towards finer phenotypic subdivisions, and the incentive for precise characterisation of the pathological processes for complex diseases.

Precision medicine is essentially the modulation of therapy and clinical management between individuals in response to their specific physiology and disease type [Collins and Varmus, 2015]. It is already a major part of many aspects of medical practice (for example,

blood typing prior to transfusion) but it has the potential to expand in scope with the increased clinical use of genomic data and -omics analysis methodology.

Incorporation of high dimensional data in medicine has several challenges. One important difficulty concerns interpretation of tests in the context of different prior probabilities of disease. Medical diagnosis is generally a Bayesian process; beginning with a prior on disease probability based on epidemiological data, a clinician makes a set of initial observations (history and examination), after which the probability distribution over potential diseases is more informative. The next stage of diagnosis generally involves choosing a set of laboratory investigations with the highest information content; hence clinicians generally observe the results of biochemical and radiological investigations in the context of an already highly-informative prior. Genetic data, by contrast, remains the same over the lifespan of a patient, and may be used at any stage in the diagnostic procedure. To be useful in making a final diagnosis, proposals for the use of genomic data in medicine must involve co-analysis with other clinical data.

The major current use of genetic data in medicine is in the diagnosis of Mendelian traits, such as Huntington's disease or cystic fibrosis. Although this is a clinically-useful application, the benefit is only to a small proportion of patients. For the large majority of the patient population, their genomic data will not be able to accurately predict whether they will develop a disease; but it may be able to inform 'ballpark' estimates of how their pattern of disease risk differs from the population average. If genomic data is to be of use to such patients, then there needs to be a way to meaningfully respond to small increases or decreases in common disease risk. An example of data of this form may be a genetic risk score of the type in chapter 6, section 6.3.5.

There are several ways in which clinical management may be slightly modulated in response to risk scores. Since risk scores can be calculated at any stage in a patient's life, one potential clinical response would be to modulate individual thresholds on clinical parameters needed to diagnose a disease. For instance, if a patient was judged by genomic data to be at a 5% increased risk of coronary artery disease, their individual threshold on HbA1C for diagnosis of type-2 diabetes could be slightly lowered, in response to a slightly increased need to maintain normoglycaemia. The storage and access of patient-specific thresholds may become more feasible with the development of electronic health record systems.

This type of intervention requires extensive comparison of results for different disease, which is the area in which my work may contribute. Predictive scores for rarer phenotypes are likely to benefit from leverage on common phenotypes, which can be facilitated using the methods in chapters 2, 3 and 6. In the above example, the intervention would only be

effective if the patient's genetic coronary disease risk was not modulated through increased genetic risk of T2D, which may require understanding the shared and distinct architectures of the two diseases; the methods developed in 5 can help direct this.

The development of precision medicine is a potentially exciting new field, and the introduction of statistical techniques for analysis of high-dimensional data will be vital to its progress. In a sense, precision medicine represents the integration of a diverse range of scientific disciplines into medical practice; some of which are discussed throughout this thesis. The further introduction of genomics into medicine, along with rigorous statistical methods and effective data management, will, I believe, come to be a milestone of the medical field.

References

[Abraham et al., 2013] Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genetic Epidemiology*, 37(2):184–195.

[Abraham et al., 2014] Abraham, G., Tye-Din, J. A., Bhalala, O. G., Kowalczyk, A., Zobel, J., and Inouye, M. (2014). Accurate and robust genomic prediction of celiac disease using statistical learning. *PLoS Genet*, 10(2):e1004137.

[Al-Matar et al., 2002] Al-Matar, M. J., Petty, R. E., Tucker, L. B., Malleson, P. N., Schroeder, M.-L., and Cabral, D. A. (2002). The early pattern of joint involvement predicts disease progression in children with oligoarticular (pauciarticular) juvenile rheumatoid arthritis. *Arthritis & Rheumatology*, 46(10):2708–2715.

[Aly et al., 2005] Aly, T. A., Ide, A., Humphrey, K., Barker, J. M., Steck, A., Erlich, H. A., Yu, L., Miao, D., Redondo, M. J., McFann, K., et al. (2005). Genetic prediction of autoimmunity: initial oligogenic prediction of anti-islet autoimmunity amongst DR3/DR4–DQ8 relatives of patients with type 1A diabetes. *Journal of autoimmunity*, 25:40–45.

[Anderson et al., 2010] Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573.

[Anderson et al., 2007] Anderson, H. R., Gupta, R., Strachan, D. P., and Limb, E. S. (2007). 50 years of asthma: UK trends from 1955 to 2004. *Thorax*, 62(1):85–90.

[Andreassen et al., 2014] Andreassen, O. A., McEvoy, L. K., Thompson, W. K., Wang, Y., Reppe, S., Schork, A. J., Zuber, V., Barrett-Connor, E., Gautvik, K., and Aukrust, P. (2014). Identifying common genetic variants in blood pressure due to polygenic pleiotropy with associated phenotypes. *Hypertension*, 63.

[Andreassen et al., 2013] Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., and Sklar, P. (2013). Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLOS Genetics*, 9(4).

[Andreasson et al., 2014] Andreasson, O. A., Harbo, H. F., Wang, Y., Thompson, W. K., Schork, A. J., Mattingsdal, M., Zuber, V., Bettella, F., and Ripke, S. (2014). Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. *Molecular psychiatry*, pages 1–8.

[Annas and Elias, 2014] Annas, G. J. and Elias, S. (2014). 23andMe and the FDA. *New England Journal of Medicine*, 370(11):985–988.

[Association et al., 1990] Association, A. S. D., Committee, D. C. S., Thorpy, M. J., et al. (1990). *The international classification of sleep disorders: diagnostic and coding manual*. American Sleep Disorders Association.

[Astle et al., 2016] Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M. A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, 167(5):1415–1429.

[Atkinson and Eisenbarth, 2001] Atkinson, M. A. and Eisenbarth, G. S. (2001). Type 1 diabetes: new perspectives on disease pathogenesis and treatment. *The Lancet*, 358(9277):221–229.

[Barrett et al., 2009] Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics*, 41(6):703–707.

[Beecham et al., 2013] Beecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., Cotsapas, C., Shahi, T. S., Spencer, C., Booth, D., Goris, A., Oturai, A., Saarela, J., and Consortium, I. M. S. G. (2013). Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature Genetics*, 45(11).

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, series B (methodological)*, 57(1):289–300.

[Bhattacharjee et al., 2012] Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., Yeager, M., Chung, C. C., Chanock, S. J., Chatterjee, N., et al. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90(5):821–835.

[Bosch et al., 2007] Bosch, X., Guilabert, A., Espinosa, G., and Mirapeix, E. (2007). Treatment of antineutrophil cytoplasmic antibody–associated vasculitis: a systematic review. *JAMA*, 298(6):655–669.

[Bottini et al., 2004] Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., MacMurray, J., Meloni, G. F., Lucarelli, P., Pellecchia, M., et al. (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type 1 diabetes. *Nature genetics*, 36(4):337.

[Bühlmann et al., 2014] Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278.

[Bulik-Sullivan et al., 2015] Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., Duncan, L., Perry, J. R., Patterson, N., Robinson, E. B., et al. (2015). An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241.

[Bullinger et al., 2004] Bullinger, L., Döhner, K., Bair, E., Fröhling, S., Schlenk, R. F., Tibshirani, R., Döhner, H., and Pollack, J. R. (2004). Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine*, 350(16):1605–1616.

[Burren et al., 2014] Burren, O. S., Guo, H., and Wallace, C. (2014). VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. *Bioinformatics*, 30(23):3342–3348.

[Canela-Xandri et al., 2017] Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2017). An atlas of genetic associations in UK Biobank. *bioRxiv*, page 176834.

[Chatterjee and Lahiri, 2011] Chatterjee, A. and Lahiri, S. (2011). Strong consistency of lasso estimators. *Sankhya A*, 73(1):55–78.

[Chatterjee and Carroll, 2005] Chatterjee, N. and Carroll, R. J. (2005). Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92(2):399–418.

[Chen et al., 2001] Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society, series B (methodological)*, 63(1):19–29.

[Cho et al., 2010] Cho, S., Kim, K., Kim, Y. J., Lee, J.-K., Cho, Y. S., Lee, J.-Y., Han, B.-G., Kim, H., Ott, J., and Park, T. (2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Annals of human genetics*, 74(5):416–428.

[Clayton, 2009] Clayton, D. G. (2009). Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet*, 5(7):e1000540.

[Colbert, 2010] Colbert, R. A. (2010). Classification of juvenile spondyloarthritis: Enthesitis-related arthritis and beyond. *Nature Reviews Rheumatology*, 6(8):477–485.

[Collins and Varmus, 2015] Collins, F. S. and Varmus, H. (2015). A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795.

[Cooper et al., 2012] Cooper, J. D., Simmonds, M. J., Walker, N. M., Burren, O., Brand, O. J., Guo, H., Wallace, C., Stevens, H., and Coleman, G. (2012). Seven newly identified loci for autoimmune thyroid disease. *Human Molecular Genetics*, 21(23):5202–5208.

[Cortes and Brown, 2011] Cortes, A. and Brown, M. A. (2011). Promise and pitfalls of the ImmunoChip. *Arthritis Research and Therapy*, 13(101).

[Cotsapas and Hafler, 2013] Cotsapas, C. and Hafler, D. A. (2013). Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends in Immunology*, 34(1):22–26.

[Cotsapas et al., 2011] Cotsapas, C., Voight, B. F., Rossin, E., Lage, K., Neale, B. M., Wallace, C., Abecasis, G. R., Barrett, J. C., and Behrens, T. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLOS Genetics*, 7(8):1–8.

[De Jager et al., 2009] De Jager, P. L., Beacher-Allan, C., Maier, L. M., Arthur, A. T., Ottoboni, L., Barcellos, L., McCauley, J. L., Sawcer, S., Goris, A., Saarela, J., Yelensky, R., Price, A., and Leppa, V. (2009). The role of the CD58 locus in multiple sclerosis. *Proceedings of the National Academy of Sciences USA*, 106(13):5264–5269.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B (methodological)*, 39(1):1–38.

[Devlin et al., 2001] Devlin, B., Roeder, K., and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theoretical Population Biology*, 60:155–166.

[Dubois et al., 2010] Dubois, P. C. A., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., Zhernakova, A., and et al, G. R. H. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nature Genetics*, 42(4):295–302.

[Duurland and Wedderburn, 2014] Duurland, C. L. and Wedderburn, L. R. (2014). Current developments in the use of biomarkers for juvenile idiopathic arthritis. *Current rheumatology reports*, 16(3):406.

[Efron et al., 2008] Efron, B. et al. (2008). Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23(1):1–22.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[Efron and Tibshirani, 2002] Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86.

[Eyre et al., 2012] Eyre, S., Bowes, J., Diogo, D., Lee, A., Barton, A., Martin, P., Zhernakova, A., Stahl, E., Viatte, S., McAllister, K., and Amos, C. I. (2012). High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nature Genetics*, 44(12):1336–1342.

[Faraco et al., 2013] Faraco, J., Lin, L., Kornum, B. R., Kenny, E. E., Trynka, G., Einen, M., Rico, T. J., Lichtner, P., Dauvilliers, Y., Arnulf, I., Lecendreux, M., Javidi, S., Geisler, P., and Mayer, G. (2013). ImmunoChip study implicates antigen presentation to T cells in narcolepsy. *PLOS Genetics*, 9(2).

[Farmer et al., 2000] Farmer, M., Petras, R. E., Hunt, L. E., Janosky, J. E., and Galandiuk, S. (2000). The importance of diagnostic accuracy in colonic inflammatory bowel disease. *The American Journal of Gastroenterology*, 95(11):3184–3188.

[Feldman and Goodrich, 1999] Feldman, R. P. and Goodrich, J. T. (1999). The Edwin Smith surgical papyrus. *Child's Nervous System*, 15(6-7):281–284.

[Ferkingstad et al., 2008] Ferkingstad, E., Frigessi, A., Rue, H., Thorleifsson, G., and Kong, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Annals of Applied Statistics*, 2(2):714–735.

[Ferrari et al., 2014] Ferrari, R., Hernandez, D. G., Nalls, M. A., Rohrer, J. D., Ramasamy, A., Kwok, J. B., Dobson-Stone, C., Brooks, W. S., Schofield, P. R., Halliday, G. M., et al. (2014). Frontotemporal dementia and its subtypes: a genome-wide association study. *The Lancet Neurology*, 13(7):686–699.

[Fortune et al., 2015] Fortune, M. D., Guo, H., Burren, O., Schofield, E., Walker, N. M., Ban, M., Sawcer, S. J., Bowes, J., Worthington, J., Barton, A., Eyre, S., Todd, J. A., and Wallace, C. (2015). Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nature Genetics*, 47:839–846.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[Fuchsberger et al., 2016] Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., et al. (2016). The genetic architecture of type 2 diabetes. *Nature*.

[Galesloot et al., 2014] Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. *PloS one*, 9(4):e95923.

[Gardner-Medwin et al., 2006] Gardner-Medwin, J. M., Killeen, O. G., Ryder, C. A., Bradshaw, K., and Johnson, K. (2006). Magnetic resonance imaging identifies features in clinically unaffected knees predicting extension of arthritis in children with monoarthritis. *The Journal of Rheumatology*, 33(11):2337–2343.

[Giambartolomei et al., 2014] Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genetics*.

[Gibson, 2012] Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13(2):135–145.

[Gioffredi et al., 2014] Gioffredi, A., Maritati, F., Oliva, E., and Buzio, C. (2014). Eosinophilic granulomatosis with polyangiitis: an overview. *Frontiers in immunology*, 5.

[Golan et al., 2014] Golan, D., Lander, E. S., and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences USA*, 111(49):5272–5281.

[Gregory et al., 2012] Gregory, A. P., Dendrou, C. A., Attfield, K. E., Haghikia, A., Xifara, D. K., Butter, F., Poschmann, G., Kaur, G., Lambert, L., Leach, O. A., et al. (2012). TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature*, 488(7412):508–511.

[Hacking, 1990] Hacking, I. (1990). *The taming of chance*, volume 17. Cambridge University Press.

[Han et al., 2016a] Han, B., Duong, D., Sul, J. H., de Bakker, P. I., Eskin, E., and Raychaudhuri, S. (2016a). A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Human Molecular Genetics*, 90.

[Han et al., 2016b] Han, B., Pouget, J. G., Slowikowski, K., Stahl, E., Lee, C. H., Diogo, D., Hu, X., Park, Y. R., Kim, E., Gregersen, P. K., et al. (2016b). A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases. *Nature genetics*, 48(7):803.

[Hasstedt et al., 2011] Hasstedt, S. J., Hanis, C. L., Das, S. K., Elbein, S. C., and The American Diabetes Association GENNID Study Group (2011). Pleiotropy of type 2 diabetes with obesity. *Journal of Human Genetics*, 56(7):491–495.

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer.

[Heeßel, 2004] Heeßel, N. P. (2004). Diagnosis, divination and disease: towards an understanding of the rationale behind the Babylonian diagnostic handbook. *Horstmanshoff and Stol*, pages 97–116.

[Hermann et al., 2015] Hermann, G., Thon, A., Mönkemöller, K., Lilienthal, E., Klinkert, C., Holder, M., Hörtenhuber, T., Vogel-Gerlicher, P., Haberland, H., Schebek, M., et al. (2015). Comorbidity of type 1 diabetes and juvenile idiopathic arthritis. *The Journal of pediatrics*, 166(4):930–935.

[Hex et al., 2012] Hex, N., Bartlett, C., Wright, D., Taylor, M., and Varley, D. (2012). Estimating the current and future costs of type 1 and type 2 diabetes in the UK, including direct health costs and indirect societal and productivity costs. *Diabetic Medicine*, 29(7):855–862.

[Hill, 1965] Hill, A. B. (1965). The environment and disease: association or causation?

[Hinks et al., 2013] Hinks, A., Cobb, J., Marion, M. C., Prahalad, S., Sudman, M., Bowes, J., Martin, P., Comeau, M. E., Sajuthi, S., Andrews, R., Brown, M., Chen, W.-M., Concannon, P., Deloukas, P., Edkins, S., Eyre, S., Gaffney, P. M., Guthery, S. L., Guthridge, J. M., Hunt, S. E., James, J. A., Keddache, M., Moser, K. L., Nigrovic, P. A., Onengut-Gumuscu, S., Onslow, M. L., Rose, C. D., Rich, S. S., Steel, K. J. A., Wakeland, E. K., Wallace, C. A., Wedderburn, L. R., Woo, P., Bohnsack, J. F., Haas, J. P., Glass, D. N., Langefeld, C. D., Thomson, W., and Thompson, S. D. (2013). Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat Genet*, 45(6):664–669.

[Hirschhorn and Daly, 2005] Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108.

[Howson et al., 2012] Howson, J. M., Cooper, J. D., Smyth, D. J., Walker, N. M., Stevens, H., She, J.-X., Eisenbarth, G. S., Rewers, M., Todd, J. A., Akolkar, B., et al. (2012). Evidence of gene-gene interaction and age-at-diagnosis effects in type 1 diabetes. *Diabetes*, 61(11):3012–3017.

[Howson et al., 2011] Howson, J. M., Rosinger, S., Smyth, D. J., Boehm, B. O., Todd, J. A., study group, A.-E., et al. (2011). Genetic analysis of adult-onset autoimmune diabetes. *Diabetes*, 60(10):2645–2653.

[Howson et al., 2009] Howson, J. M. M., Walker, N. M., Smyth, D. J., and Todd, J. A. (2009). Analysis of 19 genes for association with type 1 diabetes in the type 1 diabetes genetics consortium families. *Genes and Immunity*, 10(Suppl 1):S74–S84.

[Hyttinen et al., 2003] Hyttinen, V., Kaprio, J., Kinnunen, L., Koskenvuo, M., and Tuomilehto, J. (2003). Genetic liability of type 1 diabetes and the onset age among 22, 650 young Finnish twin pairs in a nationwide follow up study. *Diabetes*, 52(4):1052–1055.

[ImmunoBase, 2013] ImmunoBase (2013). Immunobase. www.immunobase.org.

[Javierre et al., 2016] Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., et al. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, 167(5):1369–1384.

[Jostins et al., 2012] Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P. B., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., Essers, J., Mitrociv, M., Ning, K., Cleynen, I., Theatre, E., Spain, S. L., and Raychaudhuri, S. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124.

[Klei et al., 2008] Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic epidemiology*, 32(1):9–19.

[Köhler et al., 2014] Köhler, O., Benros, M. E., Nordentoft, M., Farkouh, M. E., Iyengar, R. L., Mors, O., and Krogh, J. (2014). Effect of anti-inflammatory treatment on depression, depressive symptoms, and adverse effects: a systematic review and meta-analysis of randomized clinical trials. *JAMA psychiatry*, 71(12):1381–1391.

[Kooperberg et al., 2010] Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic epidemiology*, 34(7):643–652.

[Lane-Claypon et al., 1926] Lane-Claypon, J. E. et al. (1926). A further report on cancer of the breast with special reference to its associated antecedent conditions. *A Further Report on Cancer of the Breast with Special Reference to its Associated Antecedent Conditions.*, (32).

[Laplace, 1781] Laplace, P.-S. (1781). Mémoire sur les probabilités. *Mémoires de l'Académie Royale des sciences de Paris*, 1778:227–332.

[Lee et al., 2017] Lee, J. C., Biasci, D., Roberts, R., Gearry, R. B., Mansfield, J. C., Ahmad, T., Prescott, N. J., Satsangi, J., Wilson, D. C., Jostins, L., et al. (2017). Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nature Genetics*.

[Lee et al., 2014] Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23.

[Lee et al., 2012] Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28(19):2540–2542.

[Leslie et al., 2015] Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E. C., Cunliffe, B., 2, W. T. C. C. C., Consortium, I. M. S. G., Lawson, D. J., Falush, D., Freeman, C., Pirinen, M., Myers, S., Robinson, M., Donnelly, P., and Bodmer, W. (2015). The fine-scale genetic structure of the British population. *Nature*, 519:309–314.

[Li et al., 2015] Li, Y. R., Li, J., Zhao, S. D., Bradfield, J. P., Mentch, F. D., Maggadottir, S. M., Hou, C., Abrams, D. J., Chang, D., Gao, F., et al. (2015). Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature medicine*.

[Lichtenstein et al., 2009] Lichtenstein, P., Yip, B. H., Björk, C., Pawitan, Y., Cannon, T. D., Sullivan, P. F., and Hultman, C. M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*, 373(9659):234–239.

[Liley, 2017] Liley, J. (2017). Combining controls can improve power in two-stage association studies. *arXiv preprint:1702.02827*.

[Liley et al., 2016] Liley, J., Todd, J. A., and Wallace, C. (2016). A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. *Nature Genetics*.

[Liley and Wallace, 2015] Liley, J. and Wallace, C. (2015). A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics. *PLOS Genetics*.

[Lin and Sullivan, 2009] Lin, D. and Sullivan, P. F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *American Journal of Human Genetics*, 85(6):862–872.

[Lippert et al., 2011] Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835.

[Liu et al., 2012] Liu, J. Z., Almarri, M. A., Gaffney, D. J., Mells, G. F., Jostins, L., Cordell, H. J., Ducker, S. J., Day, D. B., Heneghan, M. A., Neuberger, J. M., Donaldson, P. T., and Bathgate, A. J. (2012). Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nature Genetics*, 44(10):1137–1141.

[Lo et al., 2015] Lo, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., and et al, N. P. (2015). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284–90.

[Lyons et al., 2017] Lyons, P., Peters, J. E., Alberici, F., Liley, J., Coulson, R. J., Astle, W., Baldini, C., Cid, M. C., Elding, H., Epplen, J., Guillevin, L., Jayne, D. R. W., Jiang, T., Gunnarsson, I., Lamprecht, P., Leslie, S., Little, M. A., Martorana, D., Moosig, F., Ohlsson, S., Ramirez, G. A., Rewerska, B., Schett, G., Sinico, R. A., Szczeklik, W., Tesar, V., Vukcevic, D., Consortium, T. E. V. G., Terrier, B., Watts, R. A., Vaglio, A., Holle, J. U., Wallace, C., and Smith, K. G. C. (2017). Genetic analysis of eosinophilic granulomatosis with polyangiitis (Churg-Strauss) informs diagnostic and therapeutic strategies. *Submitted to the Jounral of Clinical Investigation*.

[Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

[McKusick, 1969] McKusick, V. A. (1969). On lumpers and splitters, or the nosology of genetic disease. *Perspectives in biology and medicine*, 12(2):298–312.

[Moffatt et al., 2010] Moffatt, M. F., Gut, I. G., Demenais, F., Strachan, D. P., Bouzigon, E., Heath, S., Von Mutius, E., Farrall, M., Lathrop, M., and Cookson, W. O. (2010). A large-scale, consortium-based genomewide association study of asthma. *New England Journal of Medicine*, 363(13):1211–1221.

[Morris et al., 2009] Morris, A. P., Lindgren, C. M., Zeggini, E., Timpson, N. J., Frayling, T. M., Hattersley, A. T., and McCarthy, M. I. (2009). A powerful approach to subphenotype analysis in population-based genetic association studies. *Genetic Epidemiology*, 34:335–343.

[Morris et al., 2012] Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., et al. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981.

[Noble and Valdes, 2011] Noble, J. A. and Valdes, A. M. (2011). Genetics of the HLA region in the prediction of type 1 diabetes. *Current diabetes reports*, 11(6):533.

[Nyholt, 2004] Nyholt, D. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics*, 74:765–769.

[Okada et al., 2014] Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381.

[Ombrello et al., 2017] Ombrello, M. J., Arthur, V. L., Remmers, E. F., Hinks, A., Tachmazidou, I., Grom, A. A., Foell, D., Martini, A., Gattorno, M., Özen, S., et al. (2017). Genetic architecture distinguishes systemic juvenile idiopathic arthritis from other forms of juvenile idiopathic arthritis: clinical and therapeutic implications. *Annals of the rheumatic diseases*, 76(5):906–913.

[Onengut-Gumuscu et al., 2014] Onengut-Gumuscu, S., Chen, W.-M., Burren, O., Cooper, N. J., Quinlan, A. R., Mychaleckyj, J. C., Farber, E., Bonnie, J. K., Szpak, M., Schofield, E., Achuthan, P., Guo, H., and et al, M. F. (2014). Fine mapping of type 1 diabetes

susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nature Genetics*, In Press.

[O'Reilly et al., 2012] O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M.-R., and Coin, L. J. (2012). Multiphen: joint model of multiple phenotypes can increase discovery in GWAS. *PloS one*, 7(5):e34861.

[Paaby and Rockman, 2013] Paaby, A. B. and Rockman, M. V. (2013). The many faces of pleiotropy. *Trends in Genetics*, 29(2):66–73.

[Packham and Hall, 2002] Packham, J. and Hall, M. (2002). Long-term follow-up of 246 adults with juvenile idiopathic arthritis: functional outcome. *Rheumatology*, 41(12):1428–1435.

[Paneth et al., 2004] Paneth, N., Susser, E., and Susser, M. (2004). Origins and early development of the case-control study. In *A history of epidemiologic methods and concepts*, pages 291–311. Springer.

[Park et al., 2010] Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat Genet*, 42(7):570–575.

[Park and Casella, 2008] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

[Patterson and Thompson, 1971] Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, pages 545–554.

[Paynter et al., 2010] Paynter, N. P., Chasman, D. I., Paré, G., Buring, J. E., Cook, N. R., Miletich, J. P., and Ridker, P. M. (2010). Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA*, 303(7):631–637.

[Pearson and Manolio, 2008] Pearson, T. A. and Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA*, 299(11):1335–1344.

[Petty et al., 2004] Petty, R. E., Southwood, T. R., Manners, P., Baum, J., Glass, D. N., Goldenberg, J., He, X., Maldonado-Cocco, J., Orozco-Alcala, J., Prieur, A.-M., et al. (2004). International league of associations for rheumatology classification of juvenile idiopathic arthritis: second revision, edmonton, 2001. *The Journal of rheumatology*, 31(2):390.

[Piwowar et al., 2008] Piwowar, H. A., Becich, M. J., Bilofsky, H., Crowley, R. S., et al. (2008). Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med*, 5(9):e183.

[Plagnol et al., 2011] Plagnol, V., Howson, J. M. M., Smyth, D. J., Walker, N., Hafler, J. P., Wallace, C., Stevens, H., Jackson, L., Simmonds, M. J., Consortium", T. . D. G., Bingley, P. J., Gough, S. C., and Todd, J. A. (2011). Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLOS Genetics*, 7.

[Popper, 1957] Popper, K. (1957). Philosophy of science. *British Philosophy in the Mid-Century (ed. CA Mace). London: George Allen and Unwin*.

[Press, 2008] Press, D. J. (2008). Cancer of the breast—by Janet Lane-Claypon (1926): A reanalysis. Master's thesis, University of Cambridge, United Kingdom.

[Price et al., 2006] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909.

[Raab et al., 2012] Raab, A., Sengler, C., Niewerth, M., Klotsche, J., Horneff, G., Zink, A., Girschick, H., Weber, K., and Minden, K. (2012). Comorbidity profiles among adult patients with juvenile idiopathic arthritis: results of a biologic register. *Clinical and experimental rheumatology*, 31(5):796–802.

[Ramos et al., 2011] Ramos, P. S., Criswell, L. A., Moser, K. L., Comeau, M. E., Williams, A. H., Pajewski, N. M., Chung, S. A., Graham, R. R., Zidovetzki, R., Kelly, J. A., et al. (2011). A comprehensive analysis of shared loci between systemic lupus erythematosus (sle) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS Genet*, 7(12):e1002406.

[Rao, 1948] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203.

[Ravelli and Martini, 2007] Ravelli, A. and Martini, A. (2007). Juvenile idiopathic arthritis. *The Lancet*, 369(9563):767–778.

[Reményi et al., 2012] Reményi, B., Wilson, N., Steer, A., Ferreira, B., Kado, J., Kumar, K., Lawrenson, J., Maguire, G., Marijon, E., Mirabel, M., et al. (2012). World heart federation criteria for echocardiographic diagnosis of rheumatic heart disease—an evidence-based guideline. *Nature Reviews Cardiology*, 9(5):297–309.

[Ripke et al., 2014] Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K.-H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421.

[Robinson et al., 2014] Robinson, M. R., Wray, N. R., and Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30(4):124–132.

[Robinson et al., 2008] Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human heriditary disease. *American Journal of Human Genetics*, 83(5):610–615.

[Rodriguez-Calvo et al., 2016] Rodriguez-Calvo, T., Sabouri, S., Anquetil, F., and von Herrath, M. G. (2016). The viral paradigm in type 1 diabetes: Who are the main suspects? *Autoimmunity Reviews*.

[Rosenberg et al., 2010] Rosenberg, N. A., Huang, L., Jewett, E. M., Szpiech, Z. A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. *Nature reviews. Genetics*, 11(5):356.

[Sarig et al., 2012] Sarig, O., Bercovici, S., Zoller, L., Goldberg, I., Indelman, M., Nahum, S., Israeli, S., Sagiv, N., De Morentin, H. M., Katz, O., et al. (2012). Population-specific association between a polymorphic variant in ST18, encoding a pro-apoptotic molecule, and pemphigus vulgaris. *Journal of Investigative Dermatology*, 132(7):1798–1805.

[Schofield et al., 2016] Schofield, E., Carver, T., Achuthan, P., Freire-Pritchett, P., Spivakov, M., Todd, J. A., and Burren, O. S. (2016). Chicp: a web-based tool for the integrative and interactive visualization of promoter capture hi-c datasets. *Bioinformatics*, 32(16):2511–2513.

[Sehmi et al., 1992] Sehmi, R., Wardlaw, A. J., Cromwell, O., Kurihara, K., Waltmann, P., and Kay, A. B. (1992). Interleukin-5 selectively enhances the chemotactic response of eosinophils obtained from normal but not eosinophilic subjects. *Blood*, 79(11):2952–2959.

[Self and Liang, 1987] Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610.

[Shoja et al., 2011] Shoja, M. M., Rashidi, M. R., Tubbs, R. S., Etemadi, J., Abbasnejad, F., and Agutter, P. S. (2011). Legacy of Avicenna and evidence-based medicine. *International journal of cardiology*, 150(3):243–246.

[Shriner, 2012] Shriner, D. (2012). Moving toward system genetics through multiple trait analysis in genome-wide association studies. *Frontiers in genetics*, 3.

[Simon and Simon, 2007] Simon, D. and Simon, H.-U. (2007). Eosinophilic disorders. *Journal of allergy and clinical immunology*, 119(6):1291–1300.

[Simpson et al., 1945] Simpson, G. G. et al. (1945). The principles of classification and a classification of mammals. *Bull. Amer. Museum Nat. History.*, 85.

[Siroux et al., 2014] Siroux, V., González, J. R., Bouzigon, E., Curjuric, I., Boudier, A., Imboden, M., Anto, J. M., Gut, I., Jarvis, D., Lathrop, M., Omenaas, E. R., Pin, I., Wjst, M., Demenais, F., Probst-Hensch, N., Kogevinas, M., and Kauffmann, F. (2014). Genetic heterogeneity of asthma phenotypes identified by a clustering approach. *European Respiratory Journal*, 43(2):439–452.

[Sivakumaran et al., 2011] Sivakumaran, S., Agakov, F., Theodoratou, E., Theodoratou, E., Prendergast, J. G., Zgaga, L., Manolio, T. A., Rudan, I., McKeigue, P., Wilson, J. F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. *American Journal of Human Genetics*, 89:607–618.

[Skol et al., 2006] Skol, A. D., Scott, L. J., Abecasis, G. R., and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nature genetics*, 38(2):209–213.

[Smyth et al., 2008] Smyth, D., Plagnol, V., Walker, N. M., Cooper, J. D., Downes, K., Yang, J. H., Howson, J. M., Stevens, H., McManus, R., and Wijmenga, C. (2008). Shared and distinct genetic variants in type 1 diabetes and celiac disease. *New England Journal of Medicine*, 359(26):2767–2777.

[SNPedia, 2017] SNPedia (2017). SNPedia webpage; heritability. https://www.snpedia.com/index.php/Heritability. Accessed: 2017-3-30.

[Speed et al., 2012] Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, 91(6):1011–1021.

[Stahl et al., 2010] Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A. S., Zhernakova, A., Hinks, A., Guiducci, C., Chen, R., Alfredsson, L., and Amos, C. I. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, 42(6):508–516.

[Storey, 2002] Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

[Storey et al., 2003] Storey, J. D. et al. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035.

[Straumann et al., 2012] Straumann, A., Aceves, S., Blanchard, C., Collins, M., Furuta, G., Hirano, I., Schoepfer, A., Simon, D., and Simon, H.-U. (2012). Pediatric and adult eosinophilic esophagitis: similarities and differences. *Allergy*, 67(4):477–490.

[The Wellcome Trust Case Control Consortium, 2007] The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature*, 447(7145):661–678.

[Thompson et al., 2012] Thompson, S. D., Marion, M. C., Sudman, M., Ryan, M., Tsoras, M., Howard, T. D., Barnes, M. G., Ramos, P. S., Thomson, W., Hinks, A., et al. (2012). Genome-wide association analysis of juvenile idiopathic arthritis identifies a new susceptibility locus at chromosomal region 3q13. *Arthritis & Rheumatism*, 64(8):2781–2791.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

[Todd et al., 1987] Todd, J. A., Bell, J. I., and McDevitt, H. O. (1987). Hla-dq β gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature*, 329(6140):599–604.

[Traylor et al., 2013] Traylor, M., Bevan, S., Rothwell, P. M., Sudlow, C., 2, W. T. C. C. C., Dichgans, M., Markus, H. S., and Lewis, C. M. (2013). Using phenotypic heterogeneity to increase the power of genome-wide association studies: Application to age at onset of ischaemic stroke subphenotypes. *Genetic Epidemiology*, 37(5):495–503.

[Trynka et al., 2012] Trynka, G., Hunt, K. A., Bockett, N. A., Romanos, J., Mistry, V., Szperl, A., Bakker, S. F., Bardella, M. T., Bhaw-Rosun, L., Castillejo, G., de la Concha, E. G., de Almeida, R. C., Dias, K.-R. M., van Diemen, C. C., and Dubois, P. C. A. (2012). Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*, 43(12):1193–1201.

[Tsoi et al., 2012] Tsoi, L. C., Spain, S. L., Knight, J., Ellinghaus, E., Stuart, P. E., Capon, F., Ding, J., Li, Y., Tejasvi, T., Gudjonsson, J. E., Kang, H. M., and Allen, M. H. (2012). Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature Genetics*, 44(12):1341–1350.

[UK Biobank, 2007] UK Biobank (2007). UK Biobank ethics and governance framework.

[Vaglio et al., 2007] Vaglio, A., Martorana, D., Maggiore, U., Grasselli, C., Zanetti, A., Pesci, A., Garini, G., Manganelli, P., Bottero, P., Tumiati, B., et al. (2007). HLA–DRB4 as a genetic risk factor for Churg-Strauss syndrome. *Arthritis & Rheumatology*, 56(9):3159–3166.

[Wainwright, 1931] Wainwright, J. (1931). A comparison of conditions associated with breast cancer in Great Britain and America. *The American Journal of Cancer*, 15(4):2610–2645.

[Wason and Dudbridge, 2012] Wason, J. M. and Dudbridge, F. (2012). A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. *The American Journal of Human Genetics*, 90(5):760–773.

[Wasserstein and Lazar, 2016] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *Am Stat*, 70(2):129–133.

[Wen and Lu, 2013] Wen, Y. and Lu, Q. (2013). A multiclass likelihood ratio approach for genetic risk prediction allowing for phenotypic heterogeneity. *Genetic epidemiology*, 37(7):715–725.

[Wieczorek et al., 2008] Wieczorek, S., Hellmich, B., Arning, L., Moosig, F., Lamprecht, P., Gross, W., and Epplen, J. (2008). Functionally relevant variations of the interleukin-10 gene associated with antineutrophil cytoplasmic antibody–negative Churg-Strauss syndrome, but not with Wegener's granulomatosis. *Arthritis & Rheumatology*, 58(6):1839–1848.

[Wipf et al., 1999] Wipf, J. E., Lipsky, B. A., Hirschmann, J. V., Boyko, E. J., Takasugi, J., Peugeot, R. L., and Davis, C. L. (1999). Diagnosing pneumonia by physical examination: relevant or relic? *Archives of Internal Medicine*, 159(10):1082–1087.

[Wolfs et al., 2009] Wolfs, M., Hofker, M., Wijmenga, C., and Van Haeften, T. (2009). Type 2 diabetes mellitus: new genetic insights will lead to new therapeutics. *Current genomics*, 10(2):110–118.

[Wood et al., 2014] Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., , et al. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, advance online publication:–.

[Yang et al., 2011] Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G., and Hill, W. G. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43(6):519–525.

[Yarwood et al., 2015] Yarwood, A., Han, B., Raychaudhuri, S., Bowes, J., Lunt, M., Pappas, D. A., Kremer, J., Greenberg, J. D., Plenge, R., Worthington, J., et al. (2015). A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk. *Annals of the rheumatic diseases*, 74(1):170–176.

[Yu et al., 2006] Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208.

[Zablocki et al., 2014] Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M., and Thompson, W. K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics*, 30(15):2098–2104.

[Zaykin and Kozbur, 2010] Zaykin, D. V. and Kozbur, D. O. (2010). P-value based analysis for shared controls design in genome-wide association studies. *Genetic Epidemiology*, 34(7):725–738.

[Zhang et al., 2008] Zhang, H., Massey, D., Tremelling, M., and Parkes, M. (2008). Genetics of inflammatory bowel disease: clues to pathogenesis. *British medical bulletin*, 87(1):17–30.

[Zhang et al., 2010] Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355–360.

[Zhou et al., 2013] Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264.

[Zou, 2006] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Appendix A

Supplementary material for chapter 2

A.1 Supplementary note

A.1.1 Robustness of normality assumption in estimating distribution of effect sizes

The computation of cFDR with shared control groups requires an estimate of the distribution of true effect sizes η in the conditional phenotype; that is, the z values which would be observed if the MAFs in controls and cases exactly matched the true MAFs in the relevant populations. We assume that the values η are instances of a random variable H.

For any distribution of effect sizes which is non-increasing for positive effect sizes and non-decreasing for negative effect sizes, the true cFDR is less than the 'naive' cFDR obtained from applying the Andreasson formula without adjustment (theorem A.1.3, section A.1.3). That the distribution of true effect sizes should have this profile is essentially the assumption that 'small' effect sizes are more frequent than 'large' effect sizes in polygenic phenotypes, a hypothesis suggested by recent large GWAS [Wood et al., 2014].

In our algorithm, we assume the true effect sizes η follow a mixture distribution, being 0 with probability π_0 and following a normal distribution with mean 0 and standard deviation σ with probability $1-\pi_0$. This is largely for computational convenience, as it allows a simplification of a triple integral to a double (see section 2.4, equation 2.18). However, the formula can be applied for any distribution of H.

The information about the distribution of H is incorporated into the formula for $\widehat{cFDR}$ through the expression $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$. As shown in theorems 1 and A.1.3 in section A.1.3 this is larger than p_i for most reasonable distributions of H.

If the true distribution of H is different to that assumed, the estimate of $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ will be incorrect. Overestimation of $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ is a less serious problem, as it will lead to a (conservative) overestimate of $\widehat{cFDR}$. Since $\widehat{cFDR}$ is a systematically conservative estimate of $cFDR$ anyway, the expected $\widehat{cFDR}$ using the incorrect distribution of H will still provide an upper bound on the true $cFDR$ ($= Pr(H_0^{(i)} | P_i \leq p_i, P_j \leq p_j)$), although this would cause a loss in power.

If, however, an incorrect assumption on the distribution of H leads to an underestimation of $cFDR$, the estimate may no longer be conservative, meaning that the expected value of $\widehat{cFDR}$ may not be a true upper bound on $\widehat{cFDR}$. While it is the true distribution of H is generally unknown, I show in theorem A.1.4 in section A.1.3 that the distribution of H for which $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ is highest is the degenerate distribution at 0; that is, if $\pi_0 = 1$, so the worst the underestimate can be is in the case in which all SNPs are null for the conditional phenotype.

For a SNP with p values (p_i, p_j) for two phenotypes i, j, the percentage difference in $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ when using the normal approximation compared to the 'true' distribution of H is equal to the percentage difference in $\widehat{cFDR}(p_i | p_j)$ and indicates the range of potential $\widehat{cFDR}$ cutoffs for which the normal approximation will lead to a different classification of the SNP. However, this does not account for the distribution of values of (p_i, p_j); clearly some values of p_j are far more likely than others, and in general the more likely values of (p_i, p_j) tend to correspond to smaller errors. An implication of this is that although an incorrect assumption of the distribution of H could theoretically lead to a 20% over- or underestimation of the true $\widehat{cFDR}$, a much smaller error may be observed in the actual number of SNPs for which $\widehat{cFDR}$ is less than some threshold.

To show this, I simulated observed effect sizes for 10000 SNPs for a principal phenotype i and a conditional phenotype j. The Z scores for the conditional phenotype were distributed according to various non-Gaussian distributions, and the Z scores for the principal phenotype were distributed as $N(0,5)$, in order to generate values with a reasonable range. Z scores were correlated with $\rho = 0.3$, corresponding to partial sharing of controls. Values of $\widehat{cFDR}$ were computed using first the true distribution of effect sizes for the conditional phenotype and second a normal approximation. We compared the number of SNPs reaching a range of a range of thresholds for $\widehat{cFDR}$.

The first four rows of figure A.1 show the amount by which $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ is overestimated or underestimated using our technique for various distributions of H. The leftmost panels show the percentage error which arises for possible values (p_i, p_j) (p values for the principal and conditional phenotype respectively), and the middle pattern shows the

number of SNPs with $\widehat{cFDR}$ less than a given cutoff with cFDR calculated using either the true distribution or the normal approximation of H. The rightmost panel shows the true distribution of H and the normal approximation to it.

When H has a bimodal distribution, the error is almost universally below 5%. In general, if the true distribution of H follows a heavy-tailed distribution, such as the T-distribution, then because our approximation comparatively under-weights extreme H values, which would otherwise lower the estimate of $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$, in favour of values nearer zero, our approximation tends to overestimate rather than underestimate $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$, which, as discussed, is favourable. Except for very pathological distributions, the overestimation or underestimation is rarely greater than 20%. For all tested distributions, the effect on the number of SNPs with $\widehat{cFDR}$ less than a cutoff is negligible.

The lowermost panel of figure A.1 simulates a scenario in which all SNPs are null for the conditional phenotype, but in which our estimated distribution of H has parameters $pi_0 = 0.9$, $\sigma = 2$. In practice, the E-M algorithm for estimating the parameters of the distribution will generally return $\pi_0 = 1$, meaning such a scenario is unlikely to occur if the observed values of P_j do not differ substantially from their underlying distribution. This figure indicates the maximum possible underestimation of $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$, given incorrectly estimated values of π_0 and σ. Notably, it is highest when both p_i and p_j are very low.

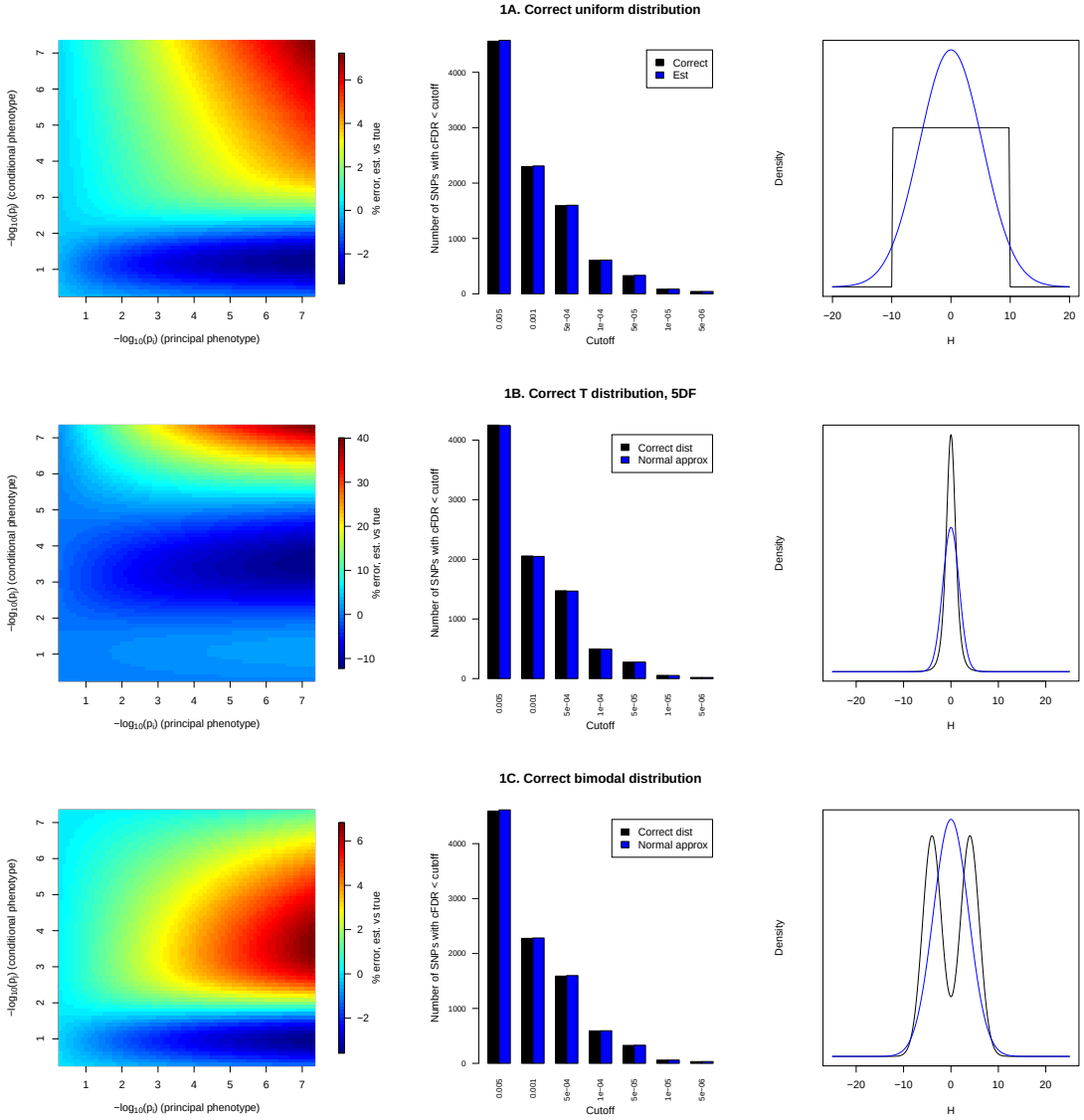
Although the left-hand panel of figure A.1 E demonstrates that there is potential for major underestimation of the expected quantile, the middle panel shows that in this scenario the mis-estimation has minimal effect on the number of SNPs reaching various $\widehat{cFDR}$ thresholds. This is because the major underestimation occurs at very low p_j values, which are very rare under the true distribution $H = 0$.

I conclude from these results that a normal approximation to the distribution of H is reasonable in most cases.

A.1.2 Estimation of the distribution of p values for the principal phenotype across SNPs null for the conditional phenotype

The distribution of p values for the principal phenotype across SNPs null for the conditional phenotype can be derived from the distribution of Z scores given in the Methods section of chapter 2.

$$(Z_i, Z_j | H_0^{(i)}) \sim \begin{cases} N(\binom{0}{0}, \binom{1\ \rho}{\rho\ 1}), & p = \pi_0^{(j)} \\ N(\binom{0}{0}, \binom{1\ \rho}{\rho\ 1+\sigma^2}), & p = 1 - \pi_0^{(j)} \end{cases} \tag{A.1}$$



Continued on next page.

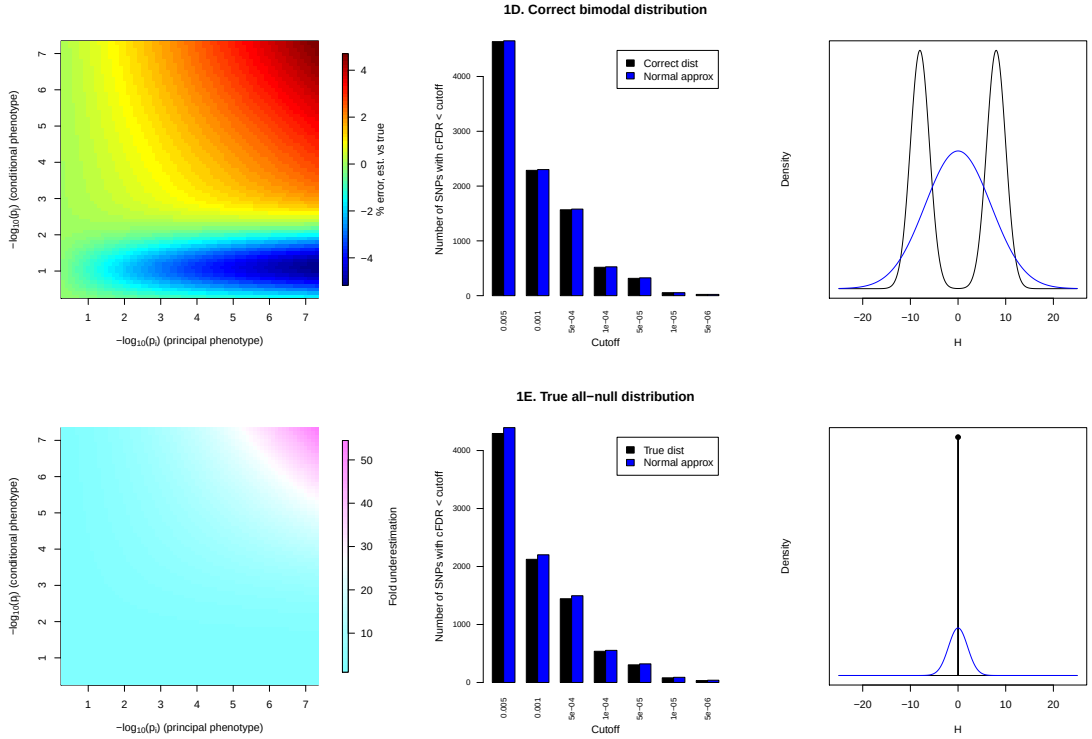


Fig. A.1 Figures showing the effect of incorrectly estimating the distribution of conditional effect sizes. If the distribution of 'true' effect sizes for the conditional phenotype (the effect sizes which would be observed in a study of equal size if the observed MAFs exactly matched population MAFs) is incorrectly assumed, the value $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ (proportional to $\widehat{cFDR}$) will be incorrectly calculated. These figures demonstrate a range of possible distributions for the true effect sizes, with an assumed value of $\pi_0 = 0.9$ in each case. In all cases, we simulated 10^5 samples from the 'true' distribution and used our E-M algorithm to find an approximating π_0 and σ. For panels A-D, the right-hand plot shows the 'true' conditional effect size distribution (uniform, T (df=5), and bimodal with near and far peaks respectively) and our approximation and the left-hand plot the percentage error in the estimate. Note that colours correspond to different values in the different graphs due to difficulties in scaling. Panel E shows the maximum fold underestimation of cFDR, which occurs the true effect sizes are all zero. The 'estimated' distribution here is not what would be estimated by our E-M algorithm (which would find the correct distribution in this case) but a simulated incorrect distribution.

While I previously computed a distribution $(Z_i, Z_j|H_0^{(i)})$ in order to compute the expected quantile, here my motivation differs. When computing the expected quantile, the motivation was to determine what the expected distribution for $(Z_i, Z_j|H_0^{(i)})$ would be *if* the SNP were null for the principal phenotype; in this section, we aim to find the expected distribution *only* for truly null SNPs. For this reason, while we fitted the earlier model to *all* values Z_j; here, I only fit the model to a subset of values Z_j which I expect to correspond to null SNPs.

I fit the model to the Z_j values for the SNPs corresponding to $P_i \geq 0.5$. Because I expect true effect sizes to be independent at SNPs which are null for the principal phenotype, this will be a representative sample of the distribution of Z_j for these SNPs. I expect that some non-null SNPs will also have $P_i \geq 0.5$; however, the proportion of such SNPs will be small, and their inclusion would be expected to lead to an estimate of the distribution of P_j biased toward low values, affecting the integral of f_0 over L more than the integral over X and making the estimate conservative.

I assume $Pr(P_j \leq p_j|P_i \geq 0.5, H_0^{(i)}) \approx Pr(P_j \leq p_j|H_0^{(i)})$ and fit π_0', σ' to the model

$$Z_j|H_0^{(i)} \sim Z_j|P_i \geq 0.5, H_0^{(i)} \sim \pi_0' N(0,1) + (1-\pi_0')N(0,\sigma'^2) \tag{A.2}$$

using expectation/maximisation.

Defining $f_{\rho,\pi_0,\sigma}(z_i,z_j)$ as

$$f_{\rho,\pi_0',\sigma'}(z_i,z_j) = \pi_0' N_{\binom{0}{0},\left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)}(z_i,z_j) + (1-\pi_0')N_{\binom{0}{0},\left(\begin{smallmatrix}1 & \rho\\ \rho & 1+\sigma'^2\end{smallmatrix}\right)}(z_i,z_j) \tag{A.3}$$

the PDF of p value pairs $(P_i,P_j) = (2\,\Phi(-|Z_i|), 2\,\Phi(-|Z_j|))$ can then be written as

$$\begin{aligned} pdf(p_i,p_j) = f_0(p_i,p_j) = \big(&f_{\rho,\pi_0',\sigma'}(-\Phi^{-1}(\frac{p_i}{2}), -\Phi^{-1}(\frac{p_j}{2}))\\ &+f_{\rho,\pi_0',\sigma'}(\Phi^{-1}(\frac{p_i}{2}), -\Phi^{-1}(\frac{p_j}{2}))\\ &+f_{\rho,\pi_0',\sigma'}(-\Phi^{-1}(\frac{p_i}{2}), \Phi^{-1}(\frac{p_j}{2}))\\ &+f_{\rho,\pi_0',\sigma'}(-\Phi^{-1}(\frac{p_i}{2}), \Phi^{-1}(\frac{p_j}{2}))\big)\\ &\times\frac{\pi}{2}\exp(\frac{1}{2}(\Phi^{-1}(\frac{p_i}{2})^2 + \Phi^{-1}(\frac{p_j}{2})^2)) \end{aligned} \tag{A.4}$$

where $\Phi(z)$ denotes as usual the normal CDF at z, and $\Phi^{-1}(p)$ its inverse at p. If the upper right vertex of M_* is on the line $P_j = 1$, then the integral of this PDF over the rectangle M^* is simply the area of M^*, as it corresponds to the area of the marginal PDF for P_i, which is $U(0,1)$ by assumption. The integral of the PDF over L is easiest to obtain by integrating

expression A.3 over the analogue of L on the plane $Z_i \times Z_j$. If $\sigma' = 1$ then the PDF is identically 1 and this integral is the area of L.

A.1.3 Overestimation of expected quantile by raw p value

Theorem 1. *Given two bivariate normal random variables* (Z_i, Z_j) *with means* $(0,0)$ *and covariance matrix* $\begin{pmatrix} 1 & \rho \\ \rho & 1+\sigma^2 \end{pmatrix}$*, define* $P_i = 2\Phi(-|Z_i|)$ *and* $P_j = 2\Phi(-|Z_j|)$ *(the p values associated with* Z_i *and* Z_j*). Then for any* p_i*,* $p_j \geq 0$

$$Pr(P_i \leq p_i | P_j \leq p_j) \geq p_i \tag{A.5}$$

with equality if and only if $\rho = 0$*,* $p_i = 0$*, or* $p_j = 0$*.*

Proof. Define $z_i = \Phi^{-1}(p_i/2)$, $z_j = \Phi^{-1}(p_j/2)$, and let

$$\begin{aligned} f_\sigma(x) &= \frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp(-\frac{1}{2(1+\sigma^2)} x^2) \\ f_\sigma(x,y,\rho) &= \frac{1}{2\pi\sqrt{1+\sigma^2-\rho^2}} \exp(-\frac{1}{2(1+\sigma^2-\rho^2)}((1+\sigma^2)x^2 - 2\rho xy + y^2)) \end{aligned} \tag{A.6}$$

the univariate and bivariate normal PDFs corresponding to the distributions of Z_j and (Z_i, Z_j) respectively.

The statement to be proved is equivalent to

$$\begin{aligned} & Pr(P_i \leq p_i, P_j \leq p_j) \geq p_i Pr(P_j \leq p_j) \\ \Leftrightarrow \quad & Pr(|Z_i| \geq z_i, |Z_j| \geq z_j) \geq Pr(|Z_i| \geq z_i) Pr(|Z_j| \geq z_j) \\ \Leftrightarrow \quad & \iint_{|x|\geq z_i, |y|\geq z_j} f_\sigma(x,y,\rho)dxdy \geq \int_{|x|\geq z_i} f_0(x)dx \int_{|y|\geq z_j} f_\sigma(y)dy \\ & \qquad = \iint_{|x|\geq z_i, |y|\geq z_j} f_\sigma(x,y,0)dxdy \end{aligned} \tag{A.7}$$

Due to the symmetry of the normal distribution and the four disjoint regions defined by $\{|x| \geq z_i, |y| \geq z_j\}$, we may rewrite both integrals over the connected region in $\{\mathbb{R}^+\}^2$ defined by $\{x \geq z_i, y \geq z_j\}$:

$$\iint_{x\geq z_i, y\geq z_j} f_\sigma(x,y,\rho) + f_\sigma(x,y,-\rho) - 2f_\sigma(x,y,0)dxdy \geq 0 \tag{A.8}$$

Rewriting $f_\sigma(x,y,\rho)$ as

$$\frac{1}{2\pi\sqrt{1+\sigma^2-\rho^2}}\exp(-\frac{1+\sigma^2}{2(1+\sigma^2-\rho^2)}(x-\frac{\rho}{1+\sigma^2})^2)\exp(-\frac{1}{2(1+\sigma^2)}y^2) \tag{A.9}$$

and noting that

$$\int_{z_i}^{\infty}\exp(-\frac{1+\sigma^2}{2(1+\sigma^2-\rho^2)}(x-\frac{\rho}{1+\sigma^2})^2 y)dx = \sqrt{2\pi\frac{(1+\sigma^2-\rho^2)}{1+\sigma^2}}\Phi(\frac{-z_i+\frac{\rho}{1+\sigma^2}y}{\sqrt{\frac{1+\sigma^2-\rho^2}{1+\sigma^2}}})$$

we can rewrite inequality A.8, after removing the common denominator $2\sqrt{2\pi(1+\sigma^2)}$, as

$$\int_{z_j}^{\infty}\big(\Phi(\frac{-z_i+\frac{\rho}{1+\sigma^2}y}{\sqrt{\frac{1+\sigma^2-\rho^2}{1+\sigma^2}}})+\Phi(\frac{-z_i-\frac{\rho}{1+\sigma^2}y}{\sqrt{\frac{1+\sigma^2-\rho^2}{1+\sigma^2}}})-2\Phi(-z_i)\big)\exp(-\frac{1}{2(1+\sigma^2)}y^2)dy>0 \tag{A.10}$$

If $z_j=0$, then the region integrated over in A.7 is two vertical strips defined by $|x|\geq z_i, y\in\mathbb{R}$. The integral of $f_\sigma(x,y,\rho)$ over this region is the integral of the marginal of Z_i over $|x|\geq z_i$, and is hence independent of ρ. Thus, in this case, the integral of $f_\sigma(x,y,\rho)+f_\sigma(x,y,-\rho)-2f_\sigma(x,y,0)$ over this region is 0. The same clearly holds if $z_i=0$ or $\rho=0$. Henceforth, we will assume $\rho>0$.

Define

$$\begin{aligned} g(y) &= \Phi(\frac{-z_i+\frac{\rho}{1+\sigma^2}y}{\sqrt{\frac{1+\sigma^2-\rho^2}{1+\sigma^2}}})+\Phi(\frac{-z_i-\frac{\rho}{1+\sigma^2}y}{\sqrt{\frac{1+\sigma^2-\rho^2}{1+\sigma^2}}})-2\Phi(-z_i) \\ h(y) &= g(y)\exp(-\frac{1}{2(1+\sigma^2)}y^2) \end{aligned} \tag{A.11}$$

Because Φ is monotonically increasing and $(1+\sigma^2-\rho^2)/(1+\sigma^2)<1$ we have $g(0)<0$. As $y\to\infty$, the first term in g tends to 1, and the second to 0. As $\Phi(-z_i)<0.5$, g therefore tends to the finite positive value $1-2\Phi(-z_i)$, and hence $h(y)$ tends to 0 from above.

We have

$$g'_{z_i}(y)=C_0\big(\exp(-\frac{(1+\sigma^2)(-\frac{\rho}{1+\sigma^2}y+z_i)^2}{2(1+\sigma^2-\rho^2)})-\exp(-\frac{(1+\sigma^2)(\frac{\rho}{1+\sigma^2}y+z_i)^2}{2(1+\sigma^2-\rho^2)})\big) \tag{A.12}$$

for a constant C_0, which can only be 0 if the exponentiated terms are equal; that is, $y=0$ or $z_i=0$. Given that h is asymptotically positive and $g(0)<0$, g is monotonically increasing

on $\mathbb{R}^+$, and is 0 at exactly one positive value y_0. Because the sign of h is the sign of g, $h(y)$ is likewise 0 if and only if $y = y_0$.

For $y > y_0$, $g(y)$ is positive, so $h(y)$ is uniformly positive across the region of integration, and hence the integral A.10 is positive for $z_j > y_0$. Because the integral of h over the positive reals is 0 we have, for any $z_j \leq y_0$

$$\begin{aligned}\int_{z_j}^{\infty} h(y)dy &= \int_{z_j}^{y_1} h(y)dy + \int_{y_1}^{\infty} h(y)dy \\ &> \int_{0}^{y_1} h(y)dy + \int_{y_1}^{\infty} h(y)dy \\ &= \int_{0}^{\infty} h(y)dy \\ &= 0 \end{aligned} \tag{A.13}$$

as required.

□

Corollary 1. *Suppose a SNP is null for two phenotypes i and j, and GWAS are performed for i and j sharing some or all controls. Let p_i be the p value obtained at the SNP for phenotype i, p_j be the p value for phenotype j, P_i, P_j the random variables from which p_i and p_j are drawn, and $H_0^{(i)}$ the null hypothesis for the SNP for phenotype i. Then p_i underestimates the probability $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$, and hence leads to an unpredictably biased and usually falsely low estimate of cFDR.*

Proof. Due to a result of Zaykin et al [Zaykin and Kozbur, 2010, Lin and Sullivan, 2009] the sharing of controls between studies induces a positive correlation between Z scores. Applying the theorem with $\sigma = 0$ yields the result. The estimated *cFDR* is computed by dividing the quantity $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ by an estimate of $Pr(P_i \leq p_i | P_j \leq p_j)$. As discussed in the methods section of chapter 2, the estimate is systematically slightly conservative (a slight overestimate) as the quantity $Pr(H_0^{(i)} | P_j \leq p_j)$ is assumed to be 1. This conservatism is lost if the estimate of $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$ is systematically low (as is the case if p_i is used as the estimator); the assertion that $E(\widehat{cFDR}) \geq \widehat{cFDR}$ no longer holds. □

Corollary 2. *Suppose a GWAS is performed for some phenotype j. For each SNP s define η_s as the Z score which would have been obtained in the GWAS if the allele frequencies for that SNP in the case and control group exactly matched the allele frequencies in the general case and control populations. Consider the values η_s as observations of a continuous random*

variable H and suppose that

$$H \sim \begin{cases} 0, & p = \pi_0^{(j)} \\ N(0,\sigma^2), & p = 1-\pi_0^{(j)} \end{cases} \tag{A.14}$$

for some π_0, σ. Suppose another GWAS is performed for some other phenotype i, sharing some or all controls with phenotype j. Let p_i and p_j be the p values obtained in these two studies for a randomly chosen SNP, and P_i, P_j the random variables associated with p_i and p_j. Define $H_0^{(i)}$ as the hypothesis that the SNP is null for phenotype i. Then p_i (equal to $Pr(P_i \leq p_i | H_0^i)$) underestimates $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$, and hence leads to an unpredictably biased and usually falsely low estimate of cFDR.

Proof. As in section 2.4.3, define

$$\begin{aligned} \Lambda_{(\rho,\sigma^2)}(z_i,z_j) &= \iint_{|x|>|z_i|,|y|>|z_j|} N_{\binom{0}{0},\left(\begin{smallmatrix}1 & \rho \\ \rho & 1+\sigma^2\end{smallmatrix}\right)}(x,y)dxdy \\ \lambda_{\sigma^2}(z_j) &= \int_{|y|>|z_j|} N_{(0,1+\sigma^2)}(y)dy \end{aligned} \tag{A.15}$$

By the results obtained in the same section, the distribution of H implies that

$$Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) = \frac{\pi_0^{(j)}\Lambda_{(\rho,0)}(z_i,z_j) + (1-\pi_0^{(j)})\Lambda_{(\rho,\sigma^2)}(z_i,z_j)}{\pi_0^{(j)}\lambda_0(z_j) + (1-\pi_0^{(j)})\lambda_{\sigma^2}(z_j)} \tag{A.16}$$

for some $\rho > 0$. From corollary 1 we have $\Lambda_{(\rho,0)}(z_i,z_j)/\lambda_0(z_j) > p_i$ and from theorem 1 we have $\Lambda_{(\rho,\sigma^2)}(z_i,z_j)/\lambda_{\sigma^2}(z_j) > p_i$. The result follows. □

Theorem 2. *Define i, j, (Z_i,Z_j), (p_i,p_j), (P_i,P_j), (p_i,p_j), η, and H as for corollary 2, but suppose H is distributed according to some unknown distribution function $k(x)$ which is non-increasing on $\mathbb{R}^+$. Then p_i underestimates $Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)})$.*

Proof. We prove the statement for $H \sim U(0,\eta_{\max})$ (the uniform distribution) for any $\eta_{\max}$. Because any decreasing function can be arbitrarily closely approximated as a linear combination of such functions with non-negative coefficients, we conclude the result.

We may assume without loss of generality that η is non-negative. Defining k again as the PDF of H, we have

$$\begin{aligned} Pr(P_i \leq p_i | P_j \leq p_j, H_0^{(i)}) &= \frac{Pr(P_i \leq p_i, P_j \leq p_j, H_0^{(i)})}{Pr(P_j \leq p_j, H_0^{(i)})} \\ &= \frac{\int_0^\infty Pr(P_i \leq p_i, P_j \leq p_j | H_0^{(i)}, H = \eta) k(\eta) d\eta}{\int_0^\infty Pr(P_j \leq p_j | H_0^{(i)}, H = \eta) k(\eta) d\eta} \end{aligned} \tag{A.17}$$

so the statement is equivalent to proving that

$$\int_0^\infty k(\eta) \big(\iint_{|x|>z_i, |y|>z_j} f_0(x, y-\eta, \rho) - f_0(x, y-\eta, 0) dx dy \big) d\eta > 0 \tag{A.18}$$

For brevity, let $q(x,y,\eta) = f_0(x, y-\eta, \rho) - f_0(x, y-\eta, 0)$. Assume that $k(\eta)$ has the form

$$k(\eta) = \begin{cases} \frac{1}{\eta_{\max}} \ \texttt{if} \ \ \eta \leq \eta_{\max} \\ 0 \ \texttt{if} \ \ \eta > \eta_{\max} \end{cases} \tag{A.19}$$

for some $\eta_{\max}$. Then inequality A.18 is equivalent to

$$\frac{1}{\eta_{\max}} \int_0^{\eta_{\max}} \big(\iint_{|x|>z_i, |y|>z_j} q(x,y,\rho) dx dy \big) d\eta \stackrel{\text{def}}{=} \frac{1}{\eta_{\max}} I(\eta_{\max}) > 0 \tag{A.20}$$

Because $f_0(x,y,\rho) = f_0(-x,-y,\rho)$ we have $q(x,y,\eta) = q(-x,\eta,y)$. Because of the symmetry of the function $q(x,y,\eta)$ about $y = \eta$, we have

$$\begin{aligned} \int_0^\infty \iint_{|x|>z_i, |y|>z_j} q(x,y,\eta) dx dy d\eta &= \frac{1}{2} \int_{-\infty}^\infty \iint_{|x|>z_i, |y|>z_j} q(x,y,\eta) dx dy d\eta \\ &= \frac{1}{2} \int_{|y|>z_j} \iint_{|x|>z_i, \eta \in \mathbb{R}} q(-x,\eta,y) dx dy d\eta \\ &= \frac{1}{2} \int_{|y|>z_j} \iint_{|x|>z_i, \eta' \in \mathbb{R}} f_0(x,\eta',\rho) - f_0(x,\eta',0) dx d\eta dy \\ &= 0 \end{aligned} \tag{A.21}$$

setting $\eta' = \eta - y$, and again using the fact that the integral over the marginal is independent of ρ. Thus $I(\eta_{\max}) \to 0$ as $\eta_{\max} \to \infty$. Again using the symmetry of the function $q(x,y,\eta)$

about $y = \eta$, we can write

$$\iint_{|x|>z_i,|y|>z_j} q(x,y,\eta)dxdy =$$

$$\begin{cases} \frac{1}{2}(\iint_{|x|>z_i,|y|>\eta+z_j} q(x,y,0)dxdy - \iint_{|x|>z_i,|y|>\eta-z_j} q(x,y,0)dxdy), & \eta > z_j \\ \frac{1}{2}(\iint_{|x|>z_i,|y|>\eta+z_j} q(x,y,0)dxdy + \iint_{|x|>z_i,|y|>z_j-\eta} q(x,y,0)dxdy), & \eta < z_j \end{cases} \tag{A.22}$$

From the second of these cases it is clear that the integrand of $I(\eta_{\max})$ is always positive if $\eta_m ax \leq z_j$. If the integrand of $I(\eta_{\max})$ (expression A.22) is to be zero, we must have

$$\begin{aligned} &\iint_{|x|>z_i,|y|>\eta+z_j} q(x,y,0)dxdy = \iint_{|x|>z_i,|y|>\eta-z_j} q(x,y,0)dxdy) \\ &\Leftrightarrow \int_{|y|>\eta+z_j}\int_{|x|>z_i} q(x,y,0)dydx = \int_{|y|>\eta-z_j}\int_{|x|>z_i} q(x,y,0)dydx \\ &\Leftrightarrow \int_{y>\eta+z_j} h(y)dy = \int_{y>\eta-z_j} h(y)dy \end{aligned} \tag{A.23}$$

using the notation $h(y)$ from theorem 1, and re-expressing the integrand over connected regions. As discussed there, the integral of $h(y)$ is negative for y less than some y_0 and positive thereafter, tending asymptotically to 0, so the integral of $h(y)$ from $y = z$ to ∞ is zero at $z = 0$, increases to a maximum at $z = x_0$, and decreases thereafter. If equality were to hold in the above, we would need the integral of $h(y)$ from z to ∞ to be equal at two values of z which are $2z_i$ apart. Because of the way the integral changes with z, the integral can only be zero for one value of η.

This implies that expression A.22 is positive for $\eta < z_j$, is zero at some value $\eta = \eta_0$, and is negative thereafter. Given that $I(\eta_{\max}) \to 0$ and is positive for $\eta_{\max} < z_j$, it must be positive for all $\eta_{\max}$. This proves the theorem for $k(\eta)$ of the form A.19.

For any general PDF $k(\eta)$ which is non-increasing on $\mathbb{R}^+$ and any $\varepsilon > 0$ we may construct $k_0(\eta)$ as a linear sum of functions k_{η_i} of the form A.19:

$$\begin{aligned} k_0(\eta) &= \sum_i w_i k_{\eta_i}(\eta) \\ &\stackrel{\text{def}}{=} \sum_i w_i \begin{cases} \frac{1}{\eta_{\max}} & \texttt{if } \eta \leq \eta_i \\ 0 & \texttt{if } \eta > \eta_i \end{cases} \end{aligned} \tag{A.24}$$

with nonnegative w_i, such that

$$\int_0^\infty |k(\eta) - k_0(\eta)| d\eta < \varepsilon \tag{A.25}$$

Now

$$\iint_{|x|>z_i,|y|>z_j} f_0(x, y-\eta, \rho) - f_0(x, y-\eta, 0) dxdy \leq 2 \iint_{|x|>z_i,|y|>z_j} f_0(x, y-\eta, \rho) dxdy \leq 2 \tag{A.26}$$

so

$$\begin{aligned} &\int_0^\infty k(\eta) \big(\iint_{|x|>z_i,|y|>z_j} f_0(x, y-\eta, \rho) - f_0(x, y-\eta, 0) dxdy \big) d\eta \\ &- \int_0^\infty k_0(\eta) \big(\iint_{|x|>z_i,|y|>z_j} f_0(x, y-\eta, \rho) - f_0(x, y-\eta, 0) dxdy \big) d\eta \\ &\leq 2 \int_0^\infty k_0(\eta) d\eta \leq 2\varepsilon \end{aligned} \tag{A.27}$$

The second term in the LHS sum above is positive for any linear combination with at least one $w_i > 0$, as it is a linear combination of integrals of the form A.20. Thus inequality A.18 holds for any k of the required form. □

Remark 1. For certain pathological distributions, it may be the case that p_i does not underestimate $Pr(P_i \leq p_i | P_j \leq p_j)$. In general, however, most GWAS are across mostly null SNPs, so in general the distribution of H (as per corollary 2) has a high mass at 0. In this case, the situation in corollary 1 is dominant, and in practice p_i is almost always an underestimate of $Pr(P_i \leq p_i | P_j \leq p_j) \leq p_i$. Except in rare cases, $p_i \neq Pr(P_i \leq p_i | P_j \leq p_j)$, and it is not reasonable to use p_i as an approximation in the current context if controls are shared between studies.

A.1.4 Maximum possible overestimation of expected quantile

Theorem 3. *Let H be a random variable taking real values, and suppose that two random variables Z_i and Z_j are distributed as*

$$(Z_i, Z_j) | H = \eta \sim N(\tbinom{0}{\eta}, \tbinom{1 \quad \rho}{\rho \quad 1+\sigma^2}) \tag{A.28}$$

that is, normally distributed with mean $(0, \eta)$ and correlation matrix $\binom{1 \quad \rho}{\rho \quad 1+\sigma^2}$. Let $P_i = 2\Phi(-Z_i)$, $P_j = 2\Phi(-Z_j)$ as usual, and (p_i, p_j) be an instance of (P_i, P_j) with $0 < p_i, p_j < 1$.

Then across all distributions of H, the value of $P(P_i \leq p_i | P_j \leq p_j)$ is maximised when H is the degenerate distribution at 0; that is, $H = 0$ with probability 1.

Proof. We will show that amongst all values of η, the value $P(P_i \leq p_i | P_j \leq p_j, H = \eta)$ is maximised when $\eta = 0$, from which the result follows.

From earlier considerations, we have

$$P(P_i \leq p_i | P_j \leq p_j, H = \eta) = \frac{\iint_{|x|>z_i, |y|>z_j} N_{\binom{0}{\eta}, \binom{1\ \rho}{\rho\ 1}}(x,y)dxdy}{\int_{|y|>z_j} N_{(\eta,1)}(y)dy}$$
$$\stackrel{\text{def}}{=} R_\eta(z_i, z_j) \tag{A.29}$$

where $z_i = -\Phi^{-1}(p_i/2)$, $z_j = -\Phi^{-1}(p_j/2)$, and N represents the normal PDF with the subscripted parameters. We may assume $\eta > 0$.

As η increases from 0, the value of $R_\eta(z_i, z_j)$ (holding (z_i, z_j) constant) tends to decrease to a minimum, then increase to an asymptote as $\eta \to \infty$ (figure A.2).

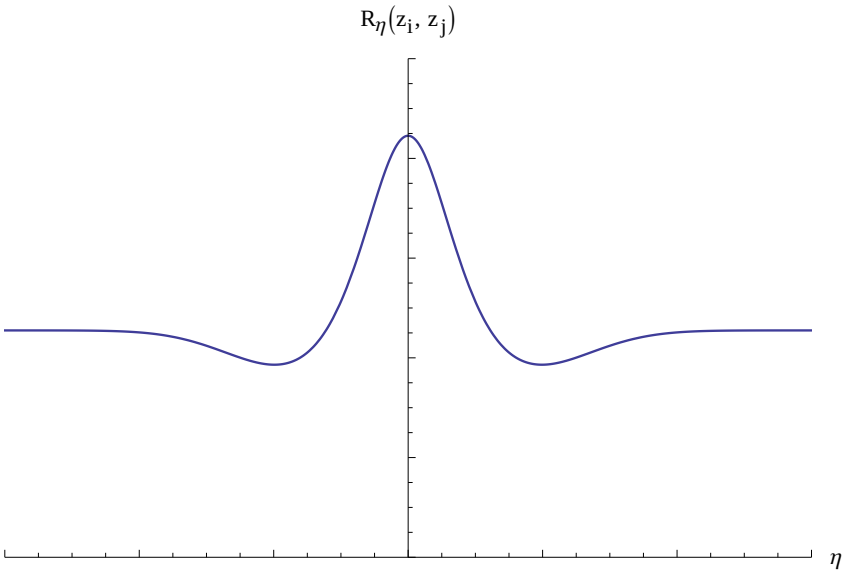


Fig. A.2 Plot of $R_\eta(z_i, z_j)$, or equivalently $Pr(P_i \leq p_i | P_j \leq p_j, H = \eta)$, as a function of η. The statement of theorem A.1.4 is that this function is maximised at $\eta = 0$, which is seen in this plot. The function decreases from 0 to a minimum then increases to an asymptote as η increases

We rewrite $R_\eta(z_i, z_j)$ as

$$R_\eta(z_i,z_j) = \frac{\int_{|x|>z_i}\int_{|y|>z_j}\frac{1}{2\pi\sqrt{(1-\rho^2)}}\exp\{(x\ y-\eta)\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}^{-1}\begin{pmatrix}x\\y-\eta\end{pmatrix}\}dydx}{\int_{|y|>z_j}\frac{1}{\sqrt{2\pi}}\exp\{-(y-\eta)^2\}dy}$$
$$= \frac{\frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{|x|>z_i}\int_{|y|>z_j}\exp\{-\frac{1}{2(1-\rho^2)}(x^2-2\rho x(y-\eta)+(y-\eta)^2)\}dydx}{\frac{1}{\sqrt{2\pi}}\int_{|y|>z_j}\exp\{-\frac{1}{2}(y-\eta)^2\}dy} \tag{A.30}$$

The numerator may be rewritten as

$$\begin{aligned}
&\frac{1}{2\pi\sqrt{(1-\rho^2)}}\int_{|x|>z_i}\int_{|y|>z_j}\exp\{-\frac{1}{2(1-\rho^2)}((y-(\eta+x\rho))^2+x^2(1-\rho^2))\}dydx\\
&=\int_{|x|>z_i}\exp\{-\frac{1}{2}x^2\}\int_{|y|>z_j}\frac{1}{2\pi\sqrt{(1-\rho^2)}}\exp\{-\frac{1}{2(1-\rho^2)}(y-(\eta+x\rho))^2\}dydx\\
&=\int_{|x|>z_i}\frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}}\big(\Phi(\frac{\eta+x\rho-z_j}{\sqrt{1-\rho^2}})+\Phi(\frac{-\eta-x\rho-z_j}{\sqrt{1-\rho^2}})\big)dx\\
&=\int_{-\infty}^{-z_i}\frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}}\big(\Phi(\frac{\eta+x\rho-z_j}{\sqrt{1-\rho^2}})+\Phi(\frac{\eta-x\rho-z_j}{\sqrt{1-\rho^2}})\\
&\qquad\qquad+\Phi(\frac{-\eta+x\rho-z_j}{\sqrt{1-\rho^2}})+\Phi(\frac{-\eta-x\rho-z_j}{\sqrt{1-\rho^2}})\big)dx\\
&\stackrel{\text{def}}{=}\int_{-\infty}^{-z_i}\frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}}K(x,z_j,\eta)dx
\end{aligned} \tag{A.31}$$

and the denominator simply as

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{|y|>z_j}\exp\{-\frac{1}{2}(y-\eta)^2\}dy &= \Phi(-z_j+\eta)+\Phi(-z_j-\eta)\\
&\stackrel{\text{def}}{=} B(z_j,\eta)
\end{aligned} \tag{A.32}$$

Since B is independent of x, we may simply include a factor of $1/B(z_j,\eta)$ in the integrand of A.31.

If $z_i = 0$, then the numerator of A.30 is an integral over the marginal of y, and hence the numerator and denominator are equal, so $R_\eta(0, z_j) = 1$. Hence

$$\begin{aligned} R_\eta(z_i, z_j) &= \int_{-\infty}^{-z_i} \frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}} \frac{K(x, z_j, \eta)}{B(z_j, \eta)} dx \\ &= 1 - \int_{-z_i}^{0} \frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}} \frac{K(x, z_j, \eta)}{B(z_j, \eta)} dx \\ &= 1 - \int_{0}^{z_i} \frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}} \frac{K(x, z_j, \eta)}{B(z_j, \eta)} dx \end{aligned}$$

given the symmetry of K in x. The problem then reduces to showing that

$$\int_0^{z_i} \frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}} \frac{K(x, z_j, \eta)}{B(z_j, \eta)} dx > \int_0^{z_i} \frac{\exp\{-\frac{1}{2}x^2\}}{\sqrt{2\pi}} \frac{K(x, z_j, 0)}{B(z_j, 0)} dx \tag{A.33}$$

for all $\eta \neq 0$.

As $z_i \to \infty$, the numerator of A.30 tends to 0 while the denominator does not change. Thus the above integral tends to 1 whatever the value of η.

For z_i approaching 0, the ratio of the RHS integral to the LHS integral tends to the ratio of the integrands at $x = 0$. This is equal to

$$\frac{K(0, z_j, \eta) B(z_j, 0)}{K(0, z_j, 0) B(z_j, \eta)} = \frac{(\Phi(\frac{\eta - z_j}{\sqrt{1-\rho^2}}) + \Phi(\frac{-\eta - z_j}{\sqrt{1-\rho^2}}))\Phi(z_j)}{(\Phi(\eta - z_j) + \Phi(-\eta + z_j))\Phi(\frac{-z_j}{\sqrt{1-\rho^2}})} \tag{A.34}$$

From the fact that Φ is convex on $\mathbb{R}^-$ and from $\Phi(x) = 1 - \Phi(-x)$, it is clear that if $b > a \geq 0$, $z > 0$ and c_1 and c_2 are such that $-z + a \leq c_1 \leq -z + b$, $-z - b \leq c_2 \leq -z + a$, then $\Phi'(c_1) > \Phi'(c_2)$. Thus $\Phi(-z + b) - \Phi(-z + a) > \Phi(-z - a) - \Phi(-z - b)$ and

$$\begin{aligned} &\Phi(\frac{-z_j + \eta}{\sqrt{1-\rho^2}}) - \Phi(\frac{-z_j}{\sqrt{1-\rho^2}} + \eta) > \Phi(\frac{-z_j}{\sqrt{1-\rho^2}} - \eta) - \Phi(\frac{-z_j - \eta}{\sqrt{1-\rho^2}}) \\ &\Leftrightarrow \Phi(\frac{-z_j + \eta}{\sqrt{1-\rho^2}}) + \Phi(\frac{-z_j - \eta}{\sqrt{1-\rho^2}}) > \Phi(\frac{-z_j}{\sqrt{1-\rho^2}} + \eta) + \Phi(\frac{-z_j}{\sqrt{1-\rho^2}} - \eta) \end{aligned} \tag{A.35}$$

We now define the function $\Phi^*(z) = \Phi(z+\eta)/\Phi(z)$, and note that the numerator of the derivative is

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z+\eta)^2}\Phi(z) - \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\Phi(z+m) \\
&\propto \int_{-\infty}^{z} e^{-\frac{1}{2}(z+\eta)^2} e^{-\frac{1}{2}x^2} - e^{-\frac{1}{2}(x+\eta)^2} e^{-\frac{1}{2}z^2} dx \\
&= \int_{-\infty}^{z} e^{-\frac{1}{2}z^2} e^{-\frac{1}{2}x^2} e^{-\frac{1}{2}\eta^2} (e^{-\eta z} - e^{-\eta x}) dx \\
&< 0
\end{aligned} \tag{A.36}$$

as the integrand is always negative. Thus

$$\begin{aligned}
&\frac{\Phi(\frac{\eta - z_j}{\sqrt{1-\rho^2}}) + \Phi(\frac{-\eta - z_j}{\sqrt{1-\rho^2}})}{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}})} \\
&> \frac{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}} + \eta) + \Phi(\frac{-z_j}{\sqrt{1-\rho^2}} - \eta)}{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}})} \\
&= \frac{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}} + \eta)}{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}})} + \frac{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}} - \eta)}{\Phi(\frac{-z_j}{\sqrt{1-\rho^2}})} \\
&> \frac{\Phi(-z_j + \eta)}{\Phi(-z_j)} + \frac{\Phi(-z_j - \eta)}{\Phi(-z_j)} \\
&= \frac{\Phi(-z_j + \eta) + \Phi(-z_j - \eta)}{\Phi(-z_j)}
\end{aligned} \tag{A.37}$$

and, rearranging, we see that expression A.34 is greater than 1. Thus for sufficiently small z_i, inequality A.33 holds.

We now show that the integrands on either side of A.33 can be equal for at most one positive value of x, which we denote x_0. Since the integrals are equal as $z_i \to \infty$, they must in fact intersect exactly once. For $z_i < x_0$ the inequality obviously holds as the LHS integrand is strictly larger than the RHS integrand. For $z_i \geq x_0$, the integral of the LHS integrand from z_i to ∞ is less than the integral of the RHS integrand over the same, so the integral of the LHS integrand from 0 to z_i is again greater than the integral of the RHS integrand from 0 to z_i. This is shown in figure A.3.

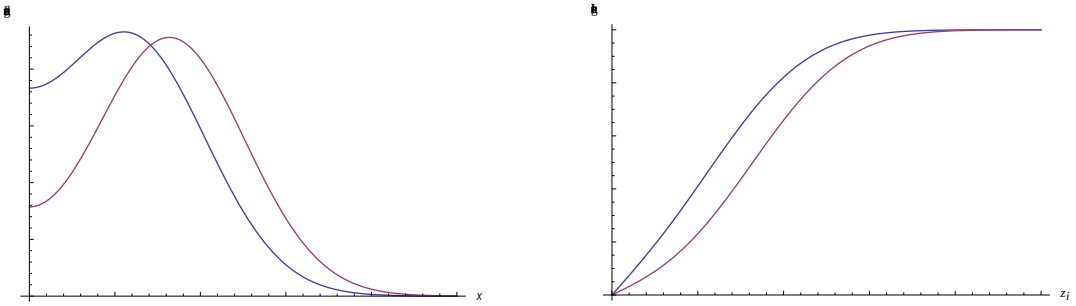


Fig. A.3 Examples of integrands (left) and integrals (right) of LHS (blue) and RHS (pink) sides of expression A.33, as functions of x and z_i respectively. We show that the integrands can only intersect once, as seen in the graph, so the integrals are never equal. This is seen on the right-hand plot.

If the integrands are equal, then

$$\frac{K(x,z_j,\eta)}{K(x,z_j,0)} = \frac{B(z_j,\eta)}{B(z_j,0)}$$

so

$$\Leftrightarrow \frac{\Phi(\frac{\eta+x\rho-z_j}{\sqrt{1-\rho^2}})+\Phi(\frac{\eta-x\rho-z_j}{\sqrt{1-\rho^2}})+\Phi(\frac{-\eta+x\rho-z_j}{\sqrt{1-\rho^2}})+\Phi(\frac{-\eta-x\rho-z_j}{\sqrt{1-\rho^2}})}{2\Phi(\frac{-x\rho-z_j}{\sqrt{1-\rho^2}})+2\Phi(\frac{x\rho-z_j}{\sqrt{1-\rho^2}})} = \frac{\Phi(\eta-z_j)+\Phi(-\eta-z_j)}{2\Phi(-z_j)} \tag{A.38}$$

For $m>0$, the function Φ is convex on $\mathbb{R}^-$ if $-x+m<0$ and the line between $(-x-m,\Phi(-x-m))$ and $(-x+m,\Phi(-x+m))$ lies strictly above $\Phi(y)$ for $y<0$. Thus we have $(\Phi(-x+m)+\Phi(-x-m))/2>\Phi(-x)$, so the RHS above is greater than 1. Considering the expression on the LHS as a function of x, we will show that it can take the required value >1 at most once on $\mathbb{R}^+$.

In analysing the LHS in isolation, the factor $\sqrt{1-\rho^2}$ merely scales η, and z_j, and the expression is to be proved for all $\eta, z_j>0$, so we may ignore the factor in this case. Likewise, the factor $\rho/\sqrt{1-\rho^2}$ simply scales x, so we may ignore it as well.

If the LHS is to take some value $1+\varepsilon/2$, then

$$\begin{aligned}&\Phi(-z_j-x+\eta)+\Phi(-z_j+x-\eta)+\Phi(-z_j+x-\eta)+\Phi(-z_j-x+\eta)\\&-(2+\varepsilon)(\Phi(-z_j-x)+\Phi(-z_j+x))=0\end{aligned} \tag{A.39}$$

Consider the function

$$L_\varepsilon(x,\eta) = (2+\varepsilon)\Phi(x) - \Phi(x-\eta) - \Phi(x+\eta) \tag{A.40}$$

so A.39 is equivalent to $L_\varepsilon(x-z_j) + L_\varepsilon(-x-z_j) = 0$. We have

$$\begin{aligned}\frac{\partial}{\partial x}L_\varepsilon(x,\eta) \stackrel{\text{def}}{=} L'_\varepsilon(x,\eta) &\propto (2+\varepsilon)e^{-\frac{1}{2}x^2} - e^{-\frac{1}{2}(x-\eta)^2} - e^{-\frac{1}{2}(x+\eta)^2} \\ &= e^{-\frac{1}{2}x^2}((2+\varepsilon) - 2e^{-\frac{1}{2}\eta^2}\cosh(\eta x))\end{aligned} \tag{A.41}$$

so $L'_\varepsilon(x,\eta)$ is zero only at only one non-negative value (and is symmetric). From the shape of the normal PDF, we have $L'_\varepsilon(0,\eta) > 0$ and because, for large enough x, we have

$$(2+\varepsilon)e^{-\frac{1}{2}x^2} < e^{-\frac{1}{2}(x-\eta)^2} \tag{A.42}$$

$L'_\varepsilon(x,\eta)$ must be asymptotically negative. Clearly $L'_\varepsilon(x,\eta) = L'_\varepsilon(-x,\eta)$, and $L'_\varepsilon(x,\eta) \to 0$ as $x \to \infty$. Furthermore, the second derivative of $L_\varepsilon(x,\eta)$ with respect to x (which we will call $L''_\varepsilon(x,\eta)$) is given by

$$\begin{aligned}L''_\varepsilon(x,\eta) &\propto (2+\varepsilon)xe^{-\frac{1}{2}x^2} - (x+\eta)e^{-\frac{1}{2}(x+\eta)^2} - (x-\eta)e^{-\frac{1}{2}(x-\eta)^2} \\ &= e^{-\frac{1}{2}x^2}((2+\varepsilon)x - e^{-\frac{1}{2}\eta^2}(2x\cosh(\eta x) - 2\eta\sinh(-\eta x)))\end{aligned} \tag{A.43}$$

The curve $y = x\cosh(\eta x) - \eta\sinh(-\eta x)$ can only intersect the line $y = x(2+\varepsilon)/(2\exp(-m^2/2))$ at one positive value of x, and given the antisymmetry of the curve and the line, the function $L''_\varepsilon(x,\eta)$ considered as a function of x has only three zeros. Given this and the earlier asymptotic properties, $L'_\varepsilon(x,\eta)$ (as a function of x) must be positive and maximal at 0, decrease monotonically to a negative minimum value, and increase monotonically thereafter. If we consider a superimposition of the curve transposed left and the curve transposed right, the transposed curves can only intersect at five points, one of which is $x = 0$.

The curves $L'_\varepsilon(x-z_j,\eta)$ and $L'_\varepsilon(-x-z_j,\eta) = L'_\varepsilon(x+z_j,\eta)$ constitute two such transposed curves and thus

$$\frac{\partial}{\partial x}(L_\varepsilon(x-z_j,\eta) + L_\varepsilon(-x-z_j,\eta)) = L'_\varepsilon(x-z_j,\eta) - L'_\varepsilon(x+z_j,\eta) \tag{A.44}$$

can have at most two positive zeros. Because $L'_\varepsilon(x)$ is asymptotically increasing as $x \to \infty$, the difference $L'_\varepsilon(x-z_j,\eta) - L'_\varepsilon(x+z_j,\eta)$ is asymptotically negative, so the derivative of

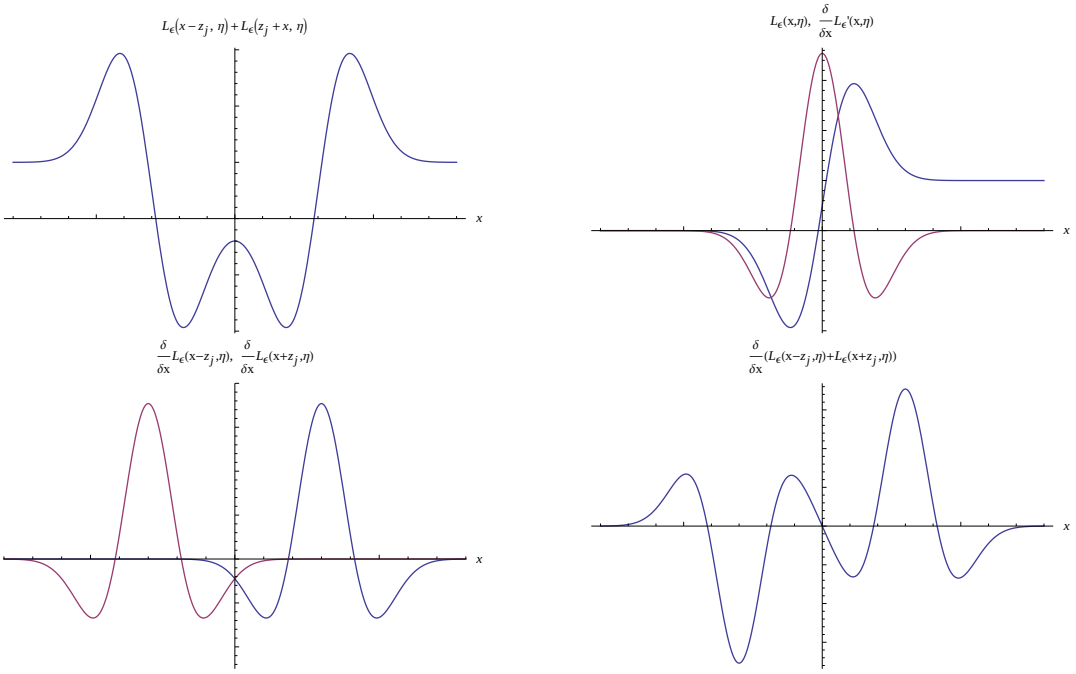


Fig. A.4 The top left plot shows $L_\varepsilon(x-z_j,\eta)+L_\varepsilon(x+z_j,\eta)$ as a function of x. We demonstrate that it is symmetric, asymptotically positive, has two stationary points for positive x, is negative at 0, and crosses zero once. These properties can be seen in this plot. The top right plot shows $L_\varepsilon(x,\eta)$ and its derivative with respect to x. The bottom right plot shows the derivative of $L_\varepsilon(x,\eta)$ transposed to the right and left by z_j. Note the limited number of points at which the two curves can cross. The bottom right plot shows the derivative with respect to x of $L_\varepsilon(x-z_j,\eta)+L_\varepsilon(x+z_j,\eta)$

$L_\varepsilon(x-z_j,\eta)+L_\varepsilon(-x-z_j,\eta)$ is asymptotically increasing as $x\to\infty$. Writing

$$L_\varepsilon(x,\eta)=2\Phi(x)-\Phi(x-\eta)-\Phi(x+\eta)+\varepsilon\Phi(x) \tag{A.45}$$

we see that as $x\to\infty$, $L_\varepsilon(x,\eta)\to\varepsilon\Phi(x)\to\varepsilon$, so $L_\varepsilon(x-z_j,\eta)+L_\varepsilon(-x-z_j,\eta)\to 2\varepsilon$. Because the derivative of $L_\varepsilon(x-z_j,\eta)+L_\varepsilon(-x-z_j,\eta)$ is asymptotically increasing toward 0, it must be asymptotically negative. Plots of L_ε, L_ε' are shown in figure A.4.

We now show that $L_\varepsilon(x-z_j,\eta)+L_\varepsilon(-x-z_j,\eta)$ is negative at $x=0$. If it were to be positive, then

$$\begin{aligned} & L_\varepsilon(-z_j,\eta)+L_\varepsilon(-z_j,\eta)>0 \\ \Leftrightarrow\ & (2+\varepsilon)\Phi(x-zj)-\Phi(-z_j-\eta)-\Phi(-z_j+\eta)>0 \\ & \Leftrightarrow 1+\frac{\varepsilon}{2}>\frac{\Phi(-z_j-\eta)+\Phi(-z_j+\eta)}{2\Phi(-zj)} \end{aligned} \tag{A.46}$$

Recalling, however, that the value of $1+\varepsilon/2$ of interest is given by the RHS of A.38, which can be rewritten as

$$\frac{\Phi(\gamma(-z_j-\eta))+\Phi(\gamma(-z_j+\eta))}{2\Phi(-\gamma zj)} \tag{A.47}$$

with $\gamma=\sqrt{1-\rho^2}<1$, we see that we must have

$$\frac{\Phi(\gamma(-z_j-\eta))+\Phi(\gamma(-z_j+\eta))}{2\Phi(-\gamma zj)} > \frac{\Phi(-z_j-\eta)+\Phi(-z_j+\eta)}{2\Phi(-zj)} \tag{A.48}$$

which is impossible as, given the shape of Φ, the LHS is strictly increasing with γ for $\gamma>0$.

Thus for the value of ε of interest, $L_\varepsilon(x-z_j,\eta)+L_\varepsilon(-x-z_j,\eta)$ is negative at $x=0$, asymptotically positive and decreasing as $x\to\infty$, and has as most two stationary points for $x>0$. Consequently, it can cross any y value between its height at 0 and its limit as $x\to\infty$ exactly once, and the result follows. □

A.2 Supplementary tables

Table A.1 Details of FDR calculation for cFDR hits

	M_{eff}	Max p	Eq uFDR	**FDR bound**
T1D	1.4	1.15×10^{-6}	6.4×10^{-5}	$\mathbf{7.32\times10^{-6}}$
ATD	1.2	1.6×10^{-6}	3.37×10^{-4}	$\mathbf{1.02\times10^{-5}}$
CEL	1.3	4.15×10^{-7}	2.44×10^{-5}	$\mathbf{5.77\times10^{-6}}$
MS	1.4	9.85×10^{-7}	7.91×10^{-5}	$\mathbf{1.19\times10^{-5}}$
NAR	3.3	7.86×10^{-8}	3.51×10^{-4}	$\mathbf{3.05\times10^{-4}}$
PBC	1.5	1.43×10^{-6}	1.71×10^{-4}	$\mathbf{2.09\times10^{-5}}$
PSO	1.9	3.11×10^{-6}	9.32×10^{-4}	$\mathbf{5.45\times10^{-5}}$
RA	2.2	2.75×10^{-6}	5.63×10^{-4}	$\mathbf{6\times10^{-5}}$
UC	1.3	6.98×10^{-7}	4.67×10^{-5}	$\mathbf{6.73\times10^{-6}}$
CRO	1.1	2.8×10^{-7}	9.9×10^{-6}	$\mathbf{2.73\times10^{-6}}$

Calculation of false discovery rates for SNPs reaching $\widehat{cFDR}$ significance levels. M_{eff} gives the 'effective number of tests', relating to the multiple testing adjustment for the multiple phenotypes conditioned upon (see Methods section for chapter 2). Max p is the maximum principal p value at which a SNP was able to be declared significant using $\widehat{cFDR}$. Eq FDR

shows the false-discovery rate we would be forced to control at in order to detect all these SNPs the principal p value alone. The FDR bound (bold) is the false discovery rate at which the set of SNPs discovered by the $\widehat{cFDR}$ method is controlled.

Table A.2 SNPs associated with T1D

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs3842727	11p15.5	2141424	8.81e-168	1.07e-115	4.9e-116	UC
rs6679677	1p13.2	114105331	4.24e-105	1.81e-100	6.1e-102	ATD
rs7396243	11p15.5	2162468	1.05e-46	1.88e-43	1.84e-43	CRO
rs3184504	12q24.12	110368991	2.67e-38	2.36e-35	1.67e-35	ATD
rs72853903	11p15.5	2155241	2e-34	1.42e-31	1.42e-31	T1D
rs61839660	10p15.1	6134703	8.42e-34	5.9e-31	5.42e-31	ATD
rs705705	12q13.2	54721771	1.19e-27	6.68e-25	5.64e-25	ATD
rs7073236	10p15.1	6146558	2.4e-20	9.09e-18	5.55e-19	RA
rs773588	1p13.2	113646347	4.68e-19	1.7e-16	7.93e-18	RA
rs12927355b	16p13.13	11102272	7.1e-19	2.56e-16	1.5e-17	MS
rs3087243	2q33.2	204447164	4.75e-18	1.63e-15	6.11e-17	RA
rs8056814	16q23.1	73809828	1.14e-16	3.54e-14	2.3e-14	MS
rs6578997	11p15.5	2176990	2.1e-16	6.39e-14	5.76e-14	ATD
rs2111485	2q24.2	162818782	1e-15	2.95e-13	8.66e-14	ATD
rs117693013	11p15.5	2155055	2e-15	5.81e-13	4.92e-13	CRO
rs6669008	1p13.2	113968084	4.79e-15	1.38e-12	5.9e-13	RA
rs7925375	11p15.5	2147731	7.52e-14	1.93e-11	1.74e-11	ATD
rs1893217	18p11.21	12799340	1.42e-13	3.57e-11	3.48e-12	RA
rs11203203	21q22.3	42709255	2.94e-13	6.94e-11	6.64e-12	ATD
rs34536443b	19p13.2	10324118	4.19e-13	9.62e-11	4.4e-12	RA
rs7068821	10q23.31	90041015	4.96e-13	1.12e-10	3.32e-11	NAR
rs72853956	11p15.5	2178847	5.76e-13	1.29e-10	8.68e-11	UC
rs1503836	1p13.2	114343021	1.74e-12	3.61e-10	1.4e-11	RA
rs516246	19q13.33	53897984	3.49e-12	6.99e-10	3.82e-10	CRO
rs72928038	6q15	91033489	4.14e-12	8.25e-10	3.11e-11	ATD
rs34843303	15q25.1	77021525	7.62e-12	1.47e-09	2.03e-10	NAR
rs3842759	11p15.5	2136445	2.23e-11	3.94e-09	3.43e-09	UC
rs72687939	1p13.2	113812829	2.83e-11	4.88e-09	3.14e-10	RA
rs7511816	1p13.2	114019237	5.12e-11	8.6e-09	3.24e-10	RA
rs2412975	22q12.2	28870590	1.35e-10	2.05e-08	1.47e-09	ATD
rs6827756	4q27	123403861	1.86e-10	2.7e-08	1.35e-08	MS

rs56994090	14q32.2	100376200	3.48e-10	4.81e-08	4.18e-08	NAR
rs4929968	11p15.5	2234264	3.63e-10	4.97e-08	2.79e-08	RA
rs7239671	18q22.2	65674240	5.19e-10	6.92e-08	9e-09	RA
rs2611215+	4q32.3	166793717	5.25e-10	6.99e-08	5.73e-08	UC
rs2304256b	19p13.2	10336652	9.48e-10	1.2e-07	6.71e-09	RA
rs151233	16p11.2	28413929	1.21e-09	1.5e-07	1.71e-08	ATD
rs736202	1p13.2	113620903	1.31e-09	1.62e-07	7.75e-09	RA
rs151181	16p11.2	28398018	5.29e-09	5.44e-07	1.3e-07	NAR
rs6043409	20p13	1564206	5.78e-09	5.82e-07	2.46e-07	NAR
rs72727394	15q14	36634314	6.72e-09	6.73e-07	5.61e-08	RA
rs193778b	16p13.13	11258712	8.12e-09	7.97e-07	1.47e-07	PBC
rs2045258	6q22.32	126716047	1.26e-08	1.18e-06	9.33e-07	CRO
rs6476839	9p24.2	4280823	1.67e-08	1.5e-06	4.54e-07	MS
rs77347828*	10p15.1	6208296	5.24e-08	4.05e-06	3.68e-06	MS
rs57582212*	10p15.1	6119041	9.91e-08	7.36e-06	1.62e-06	MS
rs12453507*	17q12	35306733	1.23e-07	8.93e-06	7.87e-07	RA
rs229533*	22q12.3	35917057	1.94e-07	1.34e-05	2.98e-06	ATD

Associated SNPs for T1D, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.3 SNPs associated with ATD

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs1023586	14q31.1	80532036	1.83e-26	5.82e-22	2.87e-22	UC
rs6679677	1p13.2	114105331	2.58e-24	1.54e-20	5.46e-22	T1D
rs11571297	2q33.2	204453248	1.38e-21	7.75e-18	5.73e-19	T1D
rs77130284*	2q33.2	204513456	3.1e-07	7.36e-05	6.69e-06	T1D
rs72928038*	6q15	91033489	4.11e-07	9.73e-05	3.2e-06	T1D
rs2030519*	3q28	189602595	7.05e-07	0.000162	5.58e-06	CEL
rs706779*	10p15.1	6138830	8.46e-07	0.000194	9.29e-06	T1D

Associated SNPs for ATD, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.4 SNPs associated with CEL

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs4686484	3q28	189601266	1.53e-47	2.04e-43	5.76e-44	ATD
rs13132308	4q27	123770564	2.6e-37	7.35e-34	5.58e-34	MS
rs6927172	6q23.3	138043868	4.29e-29	7.27e-26	2.84e-26	UC
rs76830965	3q25.33	161120372	1.61e-26	2.51e-23	7.4e-24	UC
rs1359062	1q31.2	190808095	1.39e-24	1.99e-21	1.14e-21	NAR
rs3184504	12q24.12	110368991	2.17e-20	2.03e-17	1.04e-17	MS
rs2097282	3p21.31	46353029	4.43e-20	4.05e-17	2.13e-17	CRO
rs55743914	6q22.33	128335255	3.9e-18	2.04e-15	1.5e-15	NAR
rs1250552	10q22.3	80728033	2.38e-16	1.07e-13	2.01e-14	MS
rs990171	2q12.1	102453202	3.61e-16	1.61e-13	6.97e-14	UC
rs4525910	3q25.33	161129306	5.56e-16	2.28e-13	3.83e-14	MS
rs1018326	2q31.3	181716045	8.78e-16	3.5e-13	2.42e-13	CRO
rs13003464	2p16.1	61040333	1.23e-15	4.76e-13	8.31e-14	MS
rs6441991	3p21.31	46459287	1.83e-15	6.8e-13	5.3e-13	NAR
rs1353248	3q25.33	161106253	2.2e-15	8.04e-13	3.99e-13	NAR
rs182429	6q25.3	159389562	2.36e-15	8.56e-13	1.43e-13	MS
rs1980422	2q33.2	204318641	3.9e-15	1.35e-12	3.33e-13	MS
rs77027760	6q23.3	138043754	8.3e-15	2.46e-12	5.4e-13	MS
rs61907765	11q24.3	127897147	7.93e-13	1.52e-10	5.82e-11	CRO
rs1929848+	6p21.31	35308328	1.17e-12	2.19e-10	2.09e-10	PSO
rs4445406	1p36.32	2529260	1.15e-11	1.83e-09	4.23e-10	MS
rs7104791	11q23.1	110702068	3.87e-11	5.7e-09	3.26e-09	MS
rs73001429	11q23.3	118108375	5.05e-11	7.36e-09	2.53e-09	MS
rs16896780+	6p21.31	35139530	5.99e-11	8.65e-09	7.76e-09	PSO
rs76733709	3p21.31	46256465	6.16e-11	8.86e-09	5.65e-09	UC
rs4821124	22q11.21	20309289	1.13e-10	1.55e-08	4.81e-09	CRO
rs12068671	1q24.3	170947654	2.69e-10	3.54e-08	2.26e-08	RA
rs11875687	18p11.21	12833137	3.66e-10	4.66e-08	1.52e-08	UC
rs72975916	6q22.33	128335748	6.16e-10	7.46e-08	2.27e-08	MS
rs6498114	16p13.13	10871619	1.08e-09	1.23e-07	6.12e-08	NAR
rs1050976	6p25.3	353079	3.27e-09	3.31e-07	2.82e-07	RA

rs3087243	2q33.2	204447164	4.64e-09	4.52e-07	2.79e-07	ATD
rs1893592	21q22.3	42728136	5.2e-09	5.02e-07	1.76e-07	UC
rs17602709	3q28	189553102	5.43e-09	5.22e-07	4e-07	MS
rs12142280	1q24.3	171131275	6.4e-09	6.09e-07	4.15e-07	UC
rs7616215	3p21.31	46180690	1.24e-08	1.13e-06	5.05e-07	UC
rs1107943	6q25.3	159418255	1.35e-08	1.22e-06	4.55e-07	MS
rs7162232	15q24.1	72902948	1.36e-08	1.22e-06	3.86e-07	MS
rs6715106	2q32.3	191621279	1.42e-08	1.27e-06	3.37e-07	RA
rs692890	3q25.33	161182267	1.47e-08	1.31e-06	4.52e-07	RA
rs61579022	3q13.33	120605968	1.68e-08	1.48e-06	5.65e-07	MS
rs9355697	6q25.3	159427336	2.66e-08	2.24e-06	3.18e-07	MS
rs59867199	4q27	123671681	3.11e-08	2.58e-06	5.89e-07	CRO
rs2387397	10p15.1	6430198	3.2e-08	2.66e-06	6.9e-07	RA
rs10238927	7p14.1	37400062	3.55e-08	2.92e-06	4.41e-07	MS
rs10800746	1q32.1	199148015	4.22e-08	3.43e-06	4.76e-07	MS
rs11851414*	14q24.1	68329255	7.59e-08	5.81e-06	2.37e-06	MS
rs6691768*+	1p31.3	61564451	8.56e-08	6.45e-06	2.93e-06	CRO
rs78560100*	4q27	123260921	1.13e-07	8.25e-06	1.75e-06	RA
rs6032606*+	20q13.12	44029614	1.31e-07	9.42e-06	2.89e-06	MS
rs9610686*+	22q13.1	35963797	1.9e-07	1.3e-05	3.48e-06	ATD
rs6705577*+	2p21	43212779	3.61e-07	2.17e-05	3.29e-06	MS
rs7753008*	6q15	90866360	4.15e-07	2.44e-05	3.72e-06	MS

Associated SNPs for CEL, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.5 SNPs associated with MS

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs6677309	1p13.1	116881689	7.06e-27	7.54e-22	2.64e-23	PBC
rs12928537	16p13.13	11098901	4.06e-25	8.67e-21	1.66e-22	T1D
rs2104286	10p15.1	6139051	1.63e-21	5.29e-18	1.2e-18	T1D
rs1813375	3p24.1	28053575	6.15e-17	6.57e-14	4.9e-14	PSO
rs8070345	17q23.1	55171539	4.39e-15	4.07e-12	1.82e-12	PBC
rs1800693	12p13.31	6310270	5.51e-15	5.03e-12	2.31e-12	PBC
rs41286801	1p22.1	92748052	6.25e-15	5.61e-12	1.62e-12	T1D
rs212405	6q25.3	159390547	1.09e-14	9.15e-12	6.25e-13	T1D
rs1131265	3q13.33	120705146	1.47e-14	1.19e-11	5.34e-12	PBC
rs1920296	3q13.33	123026267	4.67e-14	3.12e-11	2.49e-11	PBC
rs4780348	16p13.13	11197993	1.12e-13	6.66e-11	7.77e-12	T1D
rs9989735	2q37.1	230823698	4.66e-13	2.5e-10	2e-10	PSO
rs1359062	1q31.2	190808095	1.04e-12	5.33e-10	4.18e-11	T1D
rs11554159	19p13.11	18146944	1.42e-12	7.14e-10	3.27e-10	PBC
rs1077667	19p13.3	6619972	1.92e-12	9.31e-10	1.07e-10	T1D
rs2255214	3q13.33	123253229	2.81e-12	1.28e-09	1e-09	PSO
rs4976646	5q35.3	176721176	5.27e-12	2.28e-09	4.98e-10	T1D
rs3748817	1p36.32	2515525	6.64e-12	2.73e-09	2.62e-10	PBC
rs12087340	1p22.3	85519581	2.36e-11	8.45e-09	4.56e-09	PBC
rs17066096	6q23.3	137494601	2.69e-11	9.46e-09	5.24e-10	PBC
rs11172342	12q14.1	56474025	4.18e-11	1.44e-08	9.82e-10	PBC
rs74796499	14q31.3	87502081	3.27e-10	8.18e-08	2.4e-08	CEL
rs9282641	3q13.33	123279458	6.44e-10	1.46e-07	4.84e-08	PBC
rs6498184	16p13.13	11343491	7.58e-10	1.7e-07	8.1e-09	PBC
rs1021156	8q21.12	79738359	1.93e-09	3.98e-07	2.8e-08	PBC
rs1870071	19p13.11	16366106	1.95e-09	4.02e-07	2.34e-07	ATD
rs34383631	11q12.2	60549906	1.96e-09	4.02e-07	1.96e-07	RA
rs2288904	19p13.2	10603170	3.19e-09	6.15e-07	2.63e-07	RA
rs1014486	3q25.33	161173806	3.82e-09	7.18e-07	4.15e-08	PBC
rs706015	7p15.2	26981513	4.22e-09	7.8e-07	4.67e-07	PBC
rs2364482	12p13.31	6372392	4.56e-09	8.37e-07	1.09e-07	T1D

rs9967792	2q32.3	191682680	5.77e-09	1.02e-06	5.44e-07	CEL
rs35730213	1q32.1	199140852	6.25e-09	1.09e-06	9.25e-08	PBC
rs4410871	8q24.21	128884211	6.31e-09	1.1e-06	3.9e-07	NAR
rs11154801	6q23.3	135781048	7.41e-09	1.26e-06	1.16e-07	T1D
rs71624119	5q11.2	55476487	8.44e-09	1.4e-06	1.83e-07	T1D
rs1177209	2p16.1	60931074	1.17e-08	1.85e-06	2.03e-07	CEL
rs941816	6p21.31	36483282	1.36e-08	2.09e-06	5.24e-07	PBC
rs7923837	10q23.33	94471897	1.39e-08	2.13e-06	9.73e-07	PBC
rs11052877	12p13.31	9796957	1.61e-08	2.43e-06	2.35e-07	T1D
rs7717955	5p13.2	35898598	1.61e-08	2.43e-06	1.59e-07	PBC
rs34536443b	19p13.2	10324118	3.53e-08	4.58e-06	2.8e-07	PBC
rs11865086	16p11.2	30037994	4.93e-08	5.9e-06	6.22e-07	PBC
rs4796791*	17q21.2	37784289	5.04e-08	6.02e-06	2.62e-06	T1D
rs917116*	7p15.1	28139264	5.72e-08	6.68e-06	2.52e-06	RA
rs2445610*	8q24.21	128266270	6.75e-08	7.59e-06	2.53e-06	CEL
rs60600003*	7p14.1	37348990	6.9e-08	7.72e-06	5.86e-07	PBC
rs5884150*	7p12.2	50296113	7.87e-08	8.63e-06	3.06e-06	T1D
rs6896969*	5p13.1	40460183	8.23e-08	8.99e-06	9.24e-07	CRO
rs10892299*	11q23.3	118232053	1.07e-07	1.13e-05	1.47e-06	PBC
rs67297943*	6q23.3	138286509	1.27e-07	1.31e-05	3.63e-06	PSO
rs793108*	10p11.22	31455112	1.46e-07	1.48e-05	2.26e-06	T1D
rs2163226*	2p21	43214760	1.8e-07	1.78e-05	4.51e-06	CEL
rs7120737*	11p11.2	47658971	1.94e-07	1.9e-05	4.86e-06	CEL
rs35929052*	16q24.1	84551985	7.74e-07	6.35e-05	4.46e-06	PBC

Associated SNPs for MS, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.6 SNPs associated with NAR

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs1154155	14q11.2	22072524	2.68e-28	2.63e-23	4.22e-24	UC
rs4916312	1q25.1	171412980	7.76e-11	9.54e-07	6.89e-08	CRO
rs34843303	15q25.1	77021525	9.06e-09	5.94e-05	2.3e-06	T1D

Associated SNPs for NAR, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.7 SNPs associated with PBC

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs72678531	1p31.3	67571033	2.58e-34	2.76e-29	2.21e-30	CRO
rs574808	3q25.33	161215677	9.17e-20	2.52e-16	4.36e-17	MS
rs35188261	7q32.1	128470775	1.11e-19	2.97e-16	5.67e-17	MS
rs3024921	2q32.3	191651517	1.79e-16	2.49e-13	4.87e-14	RA
rs17122453	11q23.3	118188774	2.88e-14	3.42e-11	4.54e-12	RA
rs13092998	3q13.33	120727734	3.62e-14	4.08e-11	5.35e-12	MS
rs4728142b	7q32.1	128361203	4.62e-14	5.1e-11	2.17e-11	MS
rs1646019b	16p13.13	11267181	1.98e-13	1.83e-10	3.75e-11	MS
rs1800693	12p13.31	6310270	3.27e-13	2.89e-10	1.03e-10	MS
rs12708715	16p13.13	11085325	1.15e-12	8.29e-10	6.41e-11	MS
rs7665090	4q24	103770651	1.9e-12	1.32e-09	1.22e-10	MS
rs9303277	17q12	35229995	2.12e-12	1.46e-09	2.03e-10	MS
rs909685	22q13.1	38077617	3.37e-12	2.24e-09	4.19e-10	RA
rs6871748	5p13.2	35921739	4.57e-12	2.85e-09	2.44e-10	MS
rs1675497	3q25.33	161065076	9.82e-12	5.28e-09	6.14e-10	MS
rs34536443b	19p13.2	10324118	2.07e-11	9.5e-09	5.1e-10	MS
rs2488393	1q31.3	195988863	6.31e-11	2.55e-08	3.84e-09	MS
rs17753961	3q25.33	161142977	9.83e-11	3.7e-08	3.3e-09	MS
rs911263	14q24.1	67823346	1.04e-09	2.98e-07	2.9e-08	RA
rs6693065	1p31.3	67572606	1.18e-09	3.31e-07	6.79e-08	UC
rs11117433	16q24.1	84577017	1.11e-08	2.46e-06	1.66e-07	MS
rs17564829	17q21.31	41362429	1.62e-08	3.43e-06	3.91e-07	T1D
rs1498736	3q25.33	161177252	1.74e-08	3.63e-06	1.45e-06	CEL
rs7574865	2q32.3	191672878	2.23e-08	4.45e-06	1.11e-06	MS
rs34725611	19p13.2	10338067	2.98e-08	5.76e-06	3.48e-07	MS
rs7302763*	12q24.12	110311952	7.31e-08	1.23e-05	1.32e-06	T1D
rs1034920*+	1p13.1	116877922	1.43e-06	0.000173	3.55e-06	MS

Associated SNPs for PBC, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p

value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.8 SNPs associated with PSO

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs2474524+	10p11.21	35699658	3.76e-22	3.85e-17	8.47e-18	MS
rs76930577+	8q21.12	79722985	1.49e-18	7.65e-14	3.38e-14	NAR
rs57995211+	1q23.3	159607415	5.7e-17	1.95e-12	1.3e-12	T1D
rs78623400+	6q22.33	127278089	2.43e-16	6.22e-12	3.89e-12	T1D
rs33980500	6q21	112019955	2.5e-16	5.12e-12	4.81e-13	UC
rs143726108+	5q35.2	173175476	5.76e-16	9.84e-12	4.54e-12	UC
rs34413922+	16p11.2	28479506	6.14e-16	8.98e-12	8.98e-12	PSO
rs9525864+	13q14.11	43355611	8.28e-16	1.06e-11	1.06e-11	PSO
rs116054851+	3p21.2	50787356	1.51e-15	1.71e-11	1.71e-11	PSO
rs111251548	16p13.13	10974929	7.79e-15	7.99e-11	7.99e-11	PSO
rs114557151+	3p24.1	27794143	2.81e-14	2.62e-10	2.62e-10	PSO
rs3818818+	13q14.11	43352455	3.77e-14	3.22e-10	3.22e-10	PSO
rs13394291+	2q37.1	234166479	7.05e-14	5.55e-10	5.55e-10	PSO
rs1581803	1q21.3	150858905	4.46e-13	2.23e-09	9.28e-10	RA
rs77520588+	1p13.1	117088877	8.73e-13	3.08e-09	3.08e-09	PSO
rs17728338	5q33.1	150458511	1.05e-12	3.59e-09	2.02e-09	RA
rs10424919+	19p13.2	11697587	1.75e-12	5.61e-09	4.55e-09	NAR
rs17066690	6q23.3	138025674	2.32e-12	7.19e-09	7.19e-09	PSO
rs113866081+	6p25.3	320823	4e-12	1.05e-08	1.05e-08	PSO
rs7130650+	11q13.5	75877226	6.2e-12	1.55e-08	1.25e-08	NAR
rs34536443b	19p13.2	10324118	8.6e-12	2e-08	1.94e-09	RA
rs7722096	5q33.3	158754495	9.09e-12	2.07e-08	5.94e-09	CRO
rs456865	6q21	111739621	9.75e-12	2.17e-08	2.12e-08	RA
rs12188300	5q33.3	158762105	1.78e-11	3.8e-08	2.15e-08	UC
rs58973750+		110949512	1.06e-10	1.84e-07	1.57e-07	MS
rs117135073+	10q21.2	64025171	1.27e-10	2.13e-07	7.52e-08	UC
rs74237645+	12q24.13	111388614	4.25e-10	4.74e-07	4.04e-07	MS
rs78636848+	1q31.3	195820925	6.66e-10	7.11e-07	4.9e-07	MS
rs1416173	6q23.3	138170539	7.04e-10	7.36e-07	4.92e-07	PBC
rs114835245+	5p13.1	40548527	8.77e-10	8.9e-07	5e-07	CRO
rs57939339+	11p15.5	491841	9.87e-10	9.82e-07	4.58e-07	CRO

rs11681704+	2q37.1	234003141	1.27e-09	1.25e-06	5.73e-07	CRO
rs7865117+	9p13.3	34848595	2.89e-09	2.65e-06	2.6e-06	MS
rs6714339	2p15	61278697	5.83e-09	4.63e-06	3.3e-06	MS
rs77840275	7p14.1	37377990	8.02e-09	5.96e-06	4.47e-06	MS
rs72832931+	17q12	35246237	8.52e-09	6.24e-06	4.7e-06	MS
rs892085	19p13.2	10679092	1.05e-08	7.24e-06	6.66e-06	RA
rs30376	5q15	96146015	1.25e-08	8.03e-06	3.73e-06	UC
rs11648503+	16q22.1	67156439	1.74e-08	1.01e-05	8.43e-06	MS
rs2304256b	19p13.2	10336652	2.29e-08	1.27e-05	4.01e-06	RA
rs73246593+	8q21.12	79696198	2.3e-08	1.27e-05	2.22e-06	UC
rs57009492+	17q12	34880189	3.17e-08	1.68e-05	1.35e-05	MS
rs75430970+	11q13.5	75907370	3.62e-08	1.88e-05	8.75e-06	UC
rs75929100+	4q27	123651035	4.46e-08	2.24e-05	9.32e-06	CRO
rs11209026*	1p31.3	67478546	1.01e-07	4.8e-05	6.16e-07	UC
rs1990760*	2q24.2	162832297	2.69e-07	0.000122	1.21e-05	UC

Associated SNPs for PSO, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.9 SNPs associated with RA

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs6679677	1p13.2	114105331	3.16e-63	1.67e-58	7.89e-61	ATD
rs71624119	5q11.2	55476487	1.81e-20	3.14e-17	2.7e-18	ATD
rs34536443b	19p13.2	10324118	1.31e-13	1.05e-10	9.01e-13	T1D
rs13524	1p13.2	114029723	1.65e-11	1.2e-08	2.43e-10	T1D
rs6920220	6q23.3	138048197	1.06e-10	7.01e-08	3e-09	T1D
rs932036	4p15.2	25699960	1.15e-10	7.59e-08	2.89e-09	T1D
rs13426947	2q32.3	191641499	4.29e-10	2.32e-07	4.16e-08	T1D
rs8026898	15q23	67778471	1.36e-09	6.92e-07	9.62e-08	ATD
rs58721818	6q23.3	138285432	2.54e-09	1.25e-06	2.75e-07	T1D
rs10209110	2q11.2	100039124	7.21e-09	3.54e-06	1.43e-07	ATD
rs2812378	9p13.3	34700260	8.4e-09	4.06e-06	1.65e-06	UC
rs2228145	1q21.3	152693594	8.46e-09	4.07e-06	4.74e-07	ATD
rs2301888	1p36.13	17545317	8.77e-09	4.19e-06	2.66e-06	CRO
rs12049447	1p13.2	113724368	1.8e-08	8.29e-06	1.16e-07	T1D
rs1503836	1p13.2	114343021	3.16e-08	1.37e-05	1.65e-07	T1D
rs72685699	1p13.2	113673353	3.34e-08	1.44e-05	2.3e-07	T1D
rs39984*	5q21.1	102625191	6.29e-08	2.51e-05	7.94e-06	T1D
rs55686954*	2q33.2	204294760	8.13e-08	3.2e-05	1.96e-06	T1D
rs3087243*	2q33.2	204447164	9.27e-08	3.61e-05	3.31e-07	T1D
rs8043085*	15q14	36615432	1.12e-07	4.25e-05	1.37e-06	T1D
rs28532547*	1p36.32	2551146	1.58e-07	5.79e-05	2.81e-06	ATD
rs12936409*	17q12	35297175	2.61e-07	8.38e-05	2.59e-06	T1D
rs736202*	1p13.2	113620903	3.55e-07	0.00011	2.13e-06	T1D
rs72928038*+	6q15	91033489	5.89e-07	0.000167	2.9e-06	T1D
rs2304256b*	19p13.2	10336652	1.48e-06	0.000348	8.18e-06	T1D
rs10795791*	10p15.1	6148346	2.22e-06	0.000478	6.93e-06	T1D

Associated SNPs for RA, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p

value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.10 SNPs associated with UC

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs4654925	1p36.13	20100310	7.56e-37	8.15e-32	2.16e-32	NAR
rs7547569	1p31.3	67503956	1.56e-34	3.36e-30	2.88e-32	PSO
rs9808651	21q22.2	39388338	6.99e-28	2.79e-24	1.98e-24	PBC
rs6017342	20q13.12	42498442	3.51e-27	1.08e-23	8.64e-24	PBC
rs3024493	1q32.1	205010591	8.69e-24	2.46e-20	9.56e-21	CRO
rs10800314	1q23.3	159739413	3.1e-21	7.43e-18	3.96e-18	CRO
rs3806308+	1p36.13	20015453	1.01e-20	2.32e-17	1.35e-17	PBC
rs3197999	3p21.31	49696536	1.04e-20	2.28e-17	5.54e-18	CRO
rs10883361	10q24.2	101272958	4.21e-20	8.56e-17	1.77e-17	CRO
rs11465802	1p31.3	67458186	5.27e-20	1.05e-16	1.56e-17	CRO
rs4655215+	1p36.13	20010301	9.37e-17	7.01e-14	3.23e-14	PSO
rs12946510	17q12	35165903	4.92e-16	3.35e-13	3.88e-14	CRO
rs10758669	9p24.1	4971602	7.91e-16	5.23e-13	5.74e-14	CRO
rs11614178	12q15	66794389	1.78e-15	1.1e-12	6.66e-13	T1D
rs10781499	9q34.3	138386226	1.91e-15	1.17e-12	2.41e-13	CRO
rs12126806	1q32.1	199230448	6.27e-15	3.69e-12	3.08e-13	CRO
rs56167332	5q33.3	158760347	7.31e-14	2.52e-11	4.78e-12	CRO
rs7608910	2p16.1	61058360	3.06e-13	8.97e-11	6.65e-12	CRO
rs4366152	9q32	116604696	3.72e-13	1.08e-10	7.64e-12	CRO
rs6702254	1q32.1	205031575	7.19e-13	1.94e-10	2.38e-11	CRO
rs6466198	7q31.1	107267362	2.63e-12	6.38e-10	5.33e-10	PBC
rs1250563+	10q22.3	80717389	1.01e-11	2.16e-09	8.97e-10	CRO
rs2816958	1q32.1	198368543	1.67e-11	3.33e-09	2.46e-09	NAR
rs2823259	21q21.1	15711341	1.8e-11	3.54e-09	4.64e-10	CRO
rs17229679	2q33.1	199269002	2.55e-11	4.7e-09	3.7e-09	PBC
rs4676406	2q37.3	241227781	4.79e-11	8.35e-09	5.01e-09	CEL
rs928722	6q23.3	138015525	5.5e-11	9.49e-09	1.95e-09	CEL
rs12244675	10q21.2	64070058	6.81e-11	1.16e-08	1.95e-09	CRO
rs17085007	13q12.13	26429267	1.14e-10	1.86e-08	1.36e-08	CRO
rs4845604	1q21.3	150068304	1.29e-10	2.08e-08	5.58e-09	CRO
rs6897260	5q33.3	158735664	2.07e-09	2.45e-07	2.65e-08	CRO

rs4746475	10q21.2	64028691	2.08e-09	2.46e-07	3.7e-08	NAR
rs17229285	2q33.1	199231367	2.23e-09	2.63e-07	1.96e-07	PBC
rs115800677+	1q23.3	159635622	2.31e-09	2.71e-07	1.6e-07	CRO
rs798544	7p22.3	2729628	4.77e-09	5.19e-07	3.68e-07	PBC
rs2143178	22q13.1	37990775	4.78e-09	5.19e-07	7.93e-08	CRO
rs6062496	20q13.33	61799543	5.53e-09	5.93e-07	5.3e-08	CRO
rs7805114+	7q31.1	107237269	6.71e-09	7.08e-07	5.65e-07	PBC
rs1003643	22q13.1	38006440	7.42e-09	7.77e-07	7.4e-08	CRO
rs10891692	11q23.2	113898862	8.94e-09	9.31e-07	7.79e-07	PBC
rs11641184+	16p13.13	11612152	9.31e-09	9.67e-07	1.16e-07	CRO
rs17401847	1p36.13	20111053	9.59e-09	9.93e-07	6.01e-07	PBC
rs1893217	18p11.21	12799340	1.08e-08	1.11e-06	1.08e-07	CRO
rs3774937	4q24	103653283	1.18e-08	1.21e-06	1.01e-07	PBC
rs72684721	4q27	123197786	1.39e-08	1.4e-06	1.31e-07	CRO
rs4494327	11q13.5	75972484	2.24e-08	2.15e-06	4.84e-07	CRO
rs80243484	9q34.3	138425458	2.36e-08	2.26e-06	2.18e-07	CRO
rs12727925+	1p36.13	20014907	2.87e-08	2.71e-06	2.25e-06	PBC
rs34963268	1p36.12	22583464	3.06e-08	2.87e-06	2.08e-06	NAR
rs6088747*	20q11.22	33218265	5.69e-08	5.09e-06	2.14e-06	T1D
rs7808907*	7q32.1	128371320	1.2e-07	1.01e-05	1.5e-06	PBC
rs1479918*	4q27	123570881	1.36e-07	1.12e-05	1.86e-06	CRO
rs1790932*	18q22.2	65684380	1.38e-07	1.14e-05	3.39e-06	RA
rs79248157*+	2q32.1	185209310	1.53e-07	1.25e-05	3.55e-06	CRO
rs281423*	19p13.2	10296493	2.42e-07	1.86e-05	1.82e-06	CRO

Associated SNPs for UC, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

Table A.11 SNPs associated with CRO

RSID	Region	Pos	P val	uFDR	cFDR	CP
rs5743289	16q12.1	49314275	1.03e-86	1.11e-81	6.62e-82	CEL
rs11209026	1p31.3	67478546	1.55e-74	2.77e-70	8.62e-72	PSO
rs7517847	1p31.3	67454257	5.59e-67	2.87e-63	2.87e-63	CRO
rs6451493	5p13.1	40446692	1.1e-39	1.8e-36	7.1e-37	MS
rs11749040	5p13.1	40432182	2.02e-30	1.43e-27	6.71e-28	MS
rs3792111	2q37.1	233844429	6.24e-29	3.41e-26	2.33e-26	PSO
rs17622378	5q31.1	131806351	5.83e-27	2.35e-24	2.27e-24	PSO
rs9673419	16q12.1	49218774	1.07e-24	3.83e-22	2.94e-22	RA
rs11791262	9q34.3	138391740	9.65e-24	3.34e-21	2.81e-21	ATD
rs11236797	11q13.5	75977297	2.96e-23	1e-20	7.83e-21	RA
rs10995271	10q21.2	64108492	7.91e-23	2.63e-20	1.96e-20	RA
rs56167332	5q33.3	158760347	1.37e-21	4.27e-19	3.43e-19	RA
rs79139810+	1p31.3	67646635	1.54e-21	4.79e-19	4.06e-19	ATD
rs4957310	5p13.1	40531137	5.86e-20	1.72e-17	1.13e-17	PSO
rs4409764	10q24.2	101274227	3.64e-19	9.87e-17	9.14e-17	ATD
rs1990623+	16q12.1	49123471	3.49e-18	8.9e-16	7.91e-16	ATD
rs11597184	10p11.21	35437402	3.65e-18	9.28e-16	7.12e-16	RA
rs2143178	22q13.1	37990775	6.5e-18	1.6e-15	8.44e-16	UC
rs4958426	5q33.1	150258539	8.26e-18	2.01e-15	1.5e-15	RA
rs11646242	16q12.1	49329618	1.24e-17	2.94e-15	2.32e-15	RA
rs10758669	9p24.1	4971602	1.65e-17	3.73e-15	2.21e-15	UC
rs2823256	21q21.1	15706577	1.81e-17	4.05e-15	1.92e-15	UC
rs56735814	5q33.3	158746694	2.66e-17	5.59e-15	2.73e-15	UC
rs4643314+	16q12.1	48933456	2.36e-15	3.76e-13	3.28e-13	T1D
rs12946510	17q12	35165903	1.61e-14	2.36e-12	1.36e-12	UC
rs73818239	3p24.3	18730835	2.5e-14	3.62e-12	3.14e-12	T1D
rs9889296	17q12	29594660	4.16e-14	5.87e-12	4.24e-12	UC
rs11585473+	1p31.3	67678880	1.08e-13	1.37e-11	6.44e-12	UC
rs3197999	3p21.31	49696536	1.24e-13	1.53e-11	1.34e-11	PSO
rs6561151	13q14.11	43382706	1.3e-12	1.25e-10	9.07e-11	UC
rs10889680+	1p31.3	67541181	1.36e-12	1.3e-10	6.32e-11	UC

rs181207	16p11.2	28421031	1.53e-12	1.46e-10	9.58e-11	UC
rs7869487	9q32	116620735	3.75e-12	3.41e-10	9.81e-11	UC
rs13407913	2p23.3	24951148	3.82e-12	3.47e-10	1.91e-10	UC
rs28701841	6q21	106637023	4.08e-12	3.7e-10	2.88e-10	UC
rs34920518	18p11.21	12773086	4.49e-12	4.04e-10	1.16e-10	UC
rs55838263	1q32.1	199141351	9.52e-12	8.33e-10	2.54e-10	UC
rs2945412	17q11.2	22867770	1.44e-11	1.22e-09	1.18e-09	PSO
rs4432939	5p13.1	40706856	1.77e-11	1.46e-09	1.01e-09	UC
rs11145765	9q34.3	138523101	2.34e-11	1.89e-09	4.96e-10	UC
rs6500315+	16q12.1	49065602	3.11e-11	2.47e-09	2.15e-09	PSO
rs6062496	20q13.33	61799543	3.65e-11	2.87e-09	6.79e-10	UC
rs529866	16p13.13	11280821	5.78e-11	4.42e-09	2.33e-09	T1D
rs10191951	2p16.1	61051294	5.79e-11	4.43e-09	9.46e-10	UC
rs1558620	2q12.1	102297827	6.03e-11	4.59e-09	2.9e-09	UC
rs4807569	19p13.3	1074378	6.37e-11	4.83e-09	4.79e-09	PSO
rs1004234	5q31.1	131785000	6.95e-11	5.24e-09	3.49e-09	NAR
rs1250573	10q22.3	80712481	1.05e-10	7.62e-09	3.6e-09	T1D
rs4311543	6q21	106609325	1.08e-10	7.84e-09	6.07e-09	UC
rs780094	2p23.3	27594741	2.44e-10	1.64e-08	1.27e-08	T1D
AMBIG_3_18793384		18793384	3.15e-10	2.07e-08	4.84e-09	UC
rs118097399	6q22.33	128319926	3.48e-10	2.28e-08	9.3e-09	T1D
rs4880099	9q34.3	138530410	3.98e-10	2.58e-08	1.01e-08	UC
rs4768236	12q12	39042739	5.24e-10	3.32e-08	1.98e-08	NAR
rs925255	2p23.2	28468298	6.7e-10	4.15e-08	8.29e-09	NAR
rs5757584	22q13.1	37992496	7.74e-10	4.72e-08	1.04e-08	UC
rs1842076	5p13.1	40272775	9.97e-10	5.92e-08	2.84e-08	UC
rs9457268	6q27	167423629	1.32e-09	7.61e-08	1.66e-08	NAR
rs2284553	21q22.11	33698565	1.38e-09	7.93e-08	5.23e-08	T1D
rs3024493	1q32.1	205010591	1.43e-09	8.19e-08	2.56e-08	T1D
rs74956615	19p13.2	10288721	1.96e-09	1.09e-07	2.07e-08	PBC
rs4796793	17q21.2	37795736	2.51e-09	1.38e-07	4.58e-08	UC
rs6651252	8q24.21	129636363	2.55e-09	1.4e-07	6.1e-08	RA
rs6074022	20q13.12	44173603	3.25e-09	1.74e-07	7.39e-08	RA

rs12565884	1q32.1	199300895	4.42e-09	2.29e-07	1.38e-07	UC
rs7097656	10q23.1	82240811	5.07e-09	2.59e-07	1.26e-07	UC
rs1734907	7q22.1	100153453	5.42e-09	2.75e-07	1.62e-07	T1D
rs2074452	19p13.3	1058160	1.02e-08	4.78e-07	3.88e-07	UC
rs12041056	1p31.3	67399848	1.18e-08	5.48e-07	5.06e-07	NAR
rs2270395	16q12.1	49404333	1.21e-08	5.61e-07	5.02e-07	ATD
rs78686200	5q31.1	131550934	1.23e-08	5.7e-07	2.05e-07	UC
rs35256947	2q37.1	230869270	1.31e-08	6.04e-07	1.93e-07	T1D
rs1582515	5q33.3	158713930	2.06e-08	9.2e-07	1.82e-07	UC
rs4663340	2q37.1	233807430	2.07e-08	9.25e-07	8e-07	NAR
rs12340801+	9q31.3	113671146	2.1e-08	9.33e-07	2.8e-07	UC
rs11641016	16q24.1	84572382	2.92e-08	1.27e-06	2.75e-07	PBC
rs174535	11q12.2	61307932	3.27e-08	1.41e-06	4.7e-07	T1D
rs679574	19q13.33	53897920	3.43e-08	1.47e-06	5.5e-07	CEL
rs68143871	1q24.3	171122199	4.18e-08	1.77e-06	6.75e-07	CEL
rs72684721	4q27	123197786	4.99e-08	2.07e-06	4.27e-07	UC
rs8127691*	21q22.3	44439288	5.33e-08	2.2e-06	5.36e-07	UC
rs12924003*+	16q12.1	49637715	5.82e-08	2.38e-06	2e-06	NAR
rs10463350*	5q31.3	141514362	6.7e-08	2.69e-06	7.14e-07	UC
rs2451258*	6q25.3	159426588	8.79e-08	3.47e-06	1.03e-06	CEL
rs2476601*	1p13.2	114179091	9.49e-08	3.71e-06	3.15e-07	T1D
rs80243484*	9q34.3	138425458	1.15e-07	4.42e-06	9.04e-07	UC
rs34266232*	19p13.2	10363968	1.26e-07	4.81e-06	1.34e-06	UC
rs9358372*	6p22.3	20920567	1.38e-07	5.2e-06	2.05e-06	UC
rs79829650*	10q21.2	64050130	1.44e-07	5.4e-06	1.05e-06	NAR
rs71624119*	5q11.2	55476487	1.46e-07	5.47e-06	1.91e-06	ATD
rs12075255*	1q32.1	205028251	1.54e-07	5.77e-06	1.22e-06	UC
rs2042097*	1q31.3	195737973	2.55e-07	9.09e-06	1.58e-06	PBC

Associated SNPs for CRO, ordered by best cFDR. P values shown are after adjustment for genomic inflation. Chromosome positions are from the NCBI36 assembly. The conditional phenotype shown is the phenotype for which the cFDR was most below the relevant cutoff. Column CP is the conditional phenotype for which corrected cFDR was lowest. SNPs with p

value greater than 5×10^{-8} for the principal phenotype are asterisked, and SNP-disease associations not previously known are suffixed with a '+'

A.3 Supplementary figures

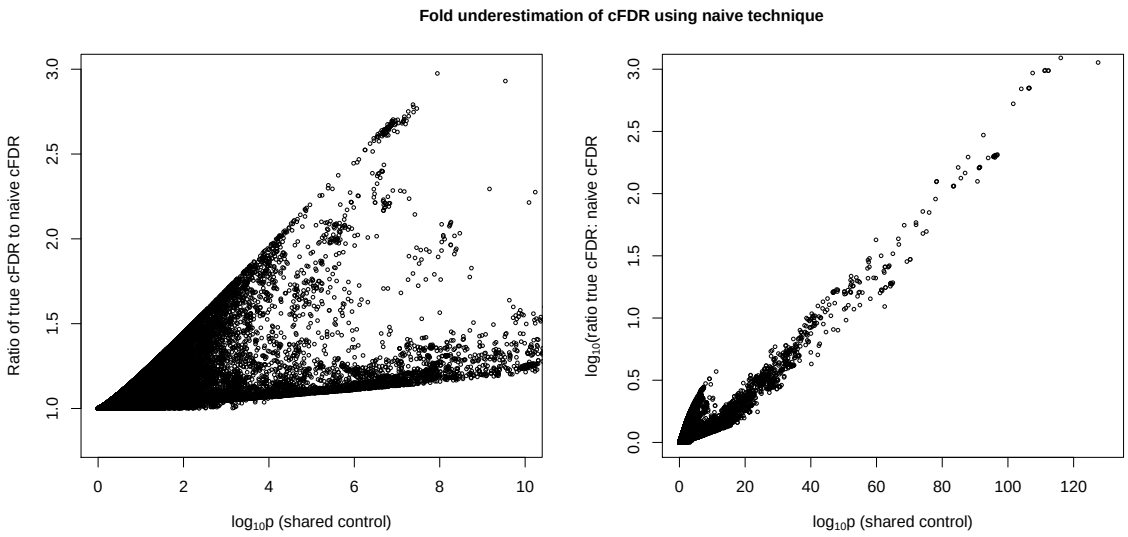


Fig. A.5 Effect of adjusting $\widehat{cFDR}$ for shared controls. This plot shows the ratio between the true $\widehat{cFDR}$ (computed using our method) to the 'naive' $\widehat{cFDR}$ (computed by naively applying the existing split-control approach to shared-control data without adjustment) for a range of p values for the principal phenotype. The p values forming the x-coordinates were obtained from the shared-control design.

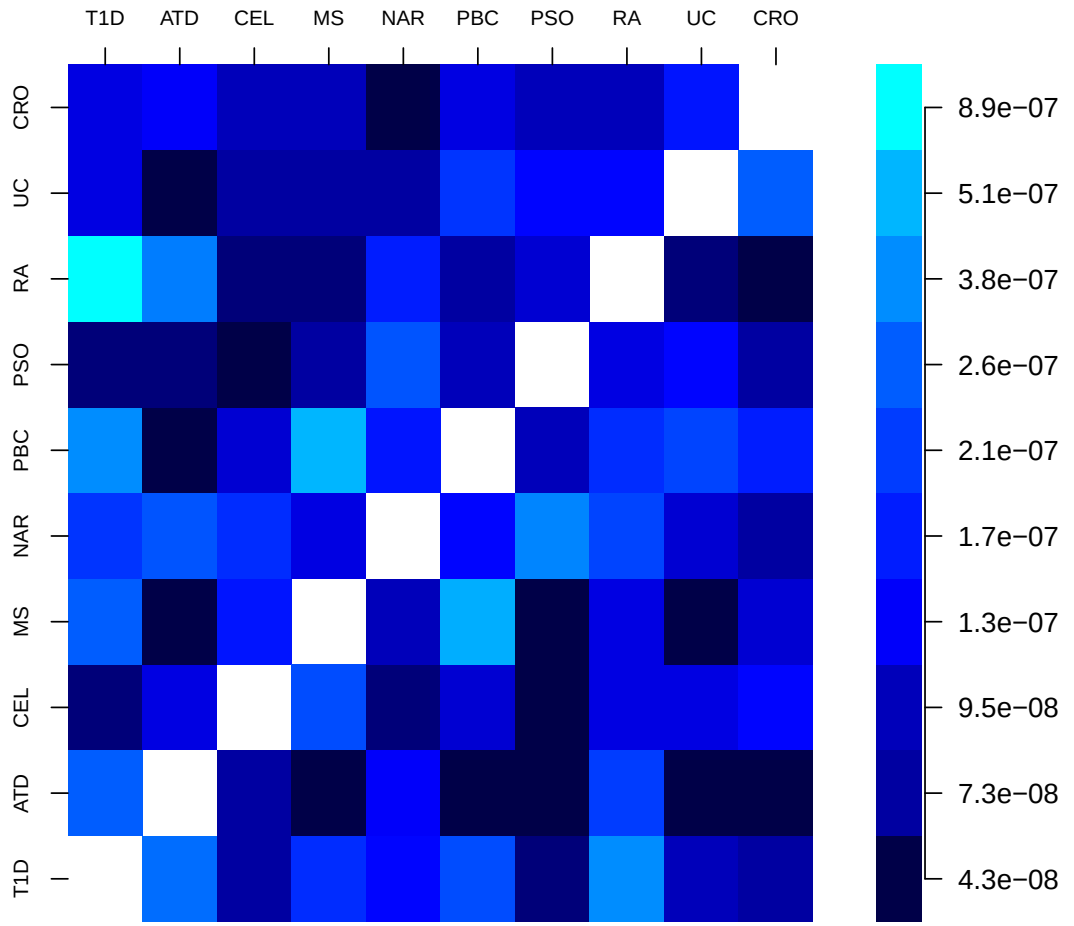


Fig. A.6 Summary of pleiotropy between phenotypes. The colour for phenotype i (horizontal) and phenotype j (vertical) corresponds to the p-value cutoff for significance for phenotype i, given that a p-value cutoff for phenotype j is less than 5×10^{-6}.

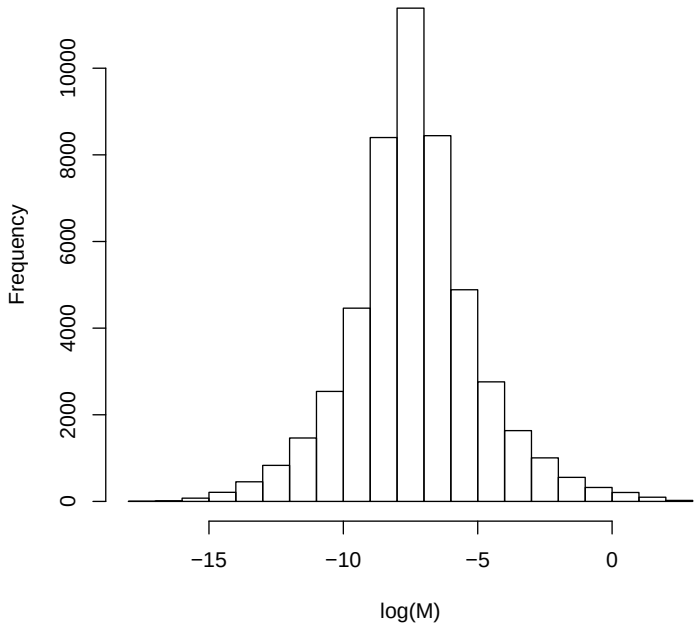
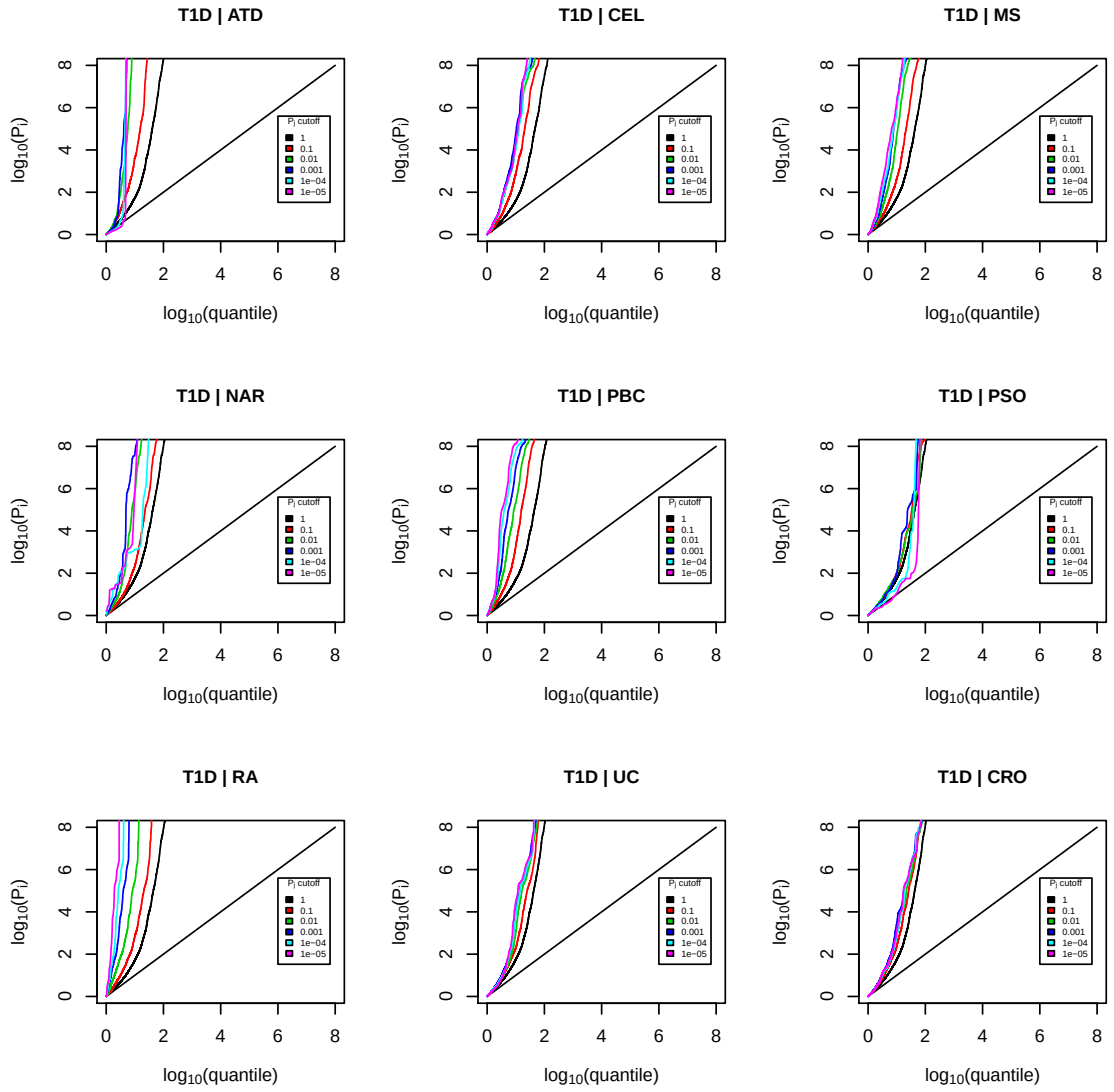
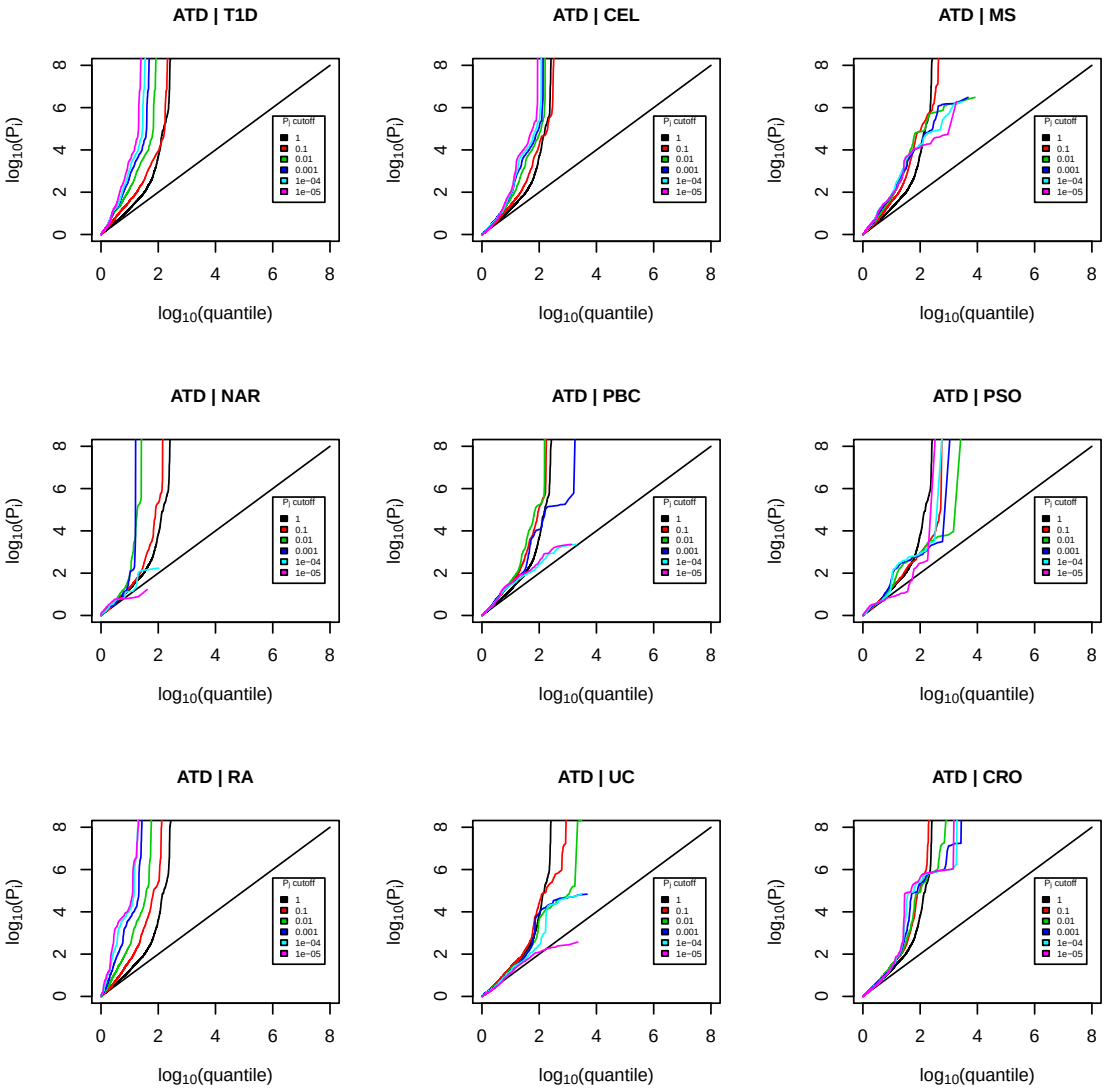
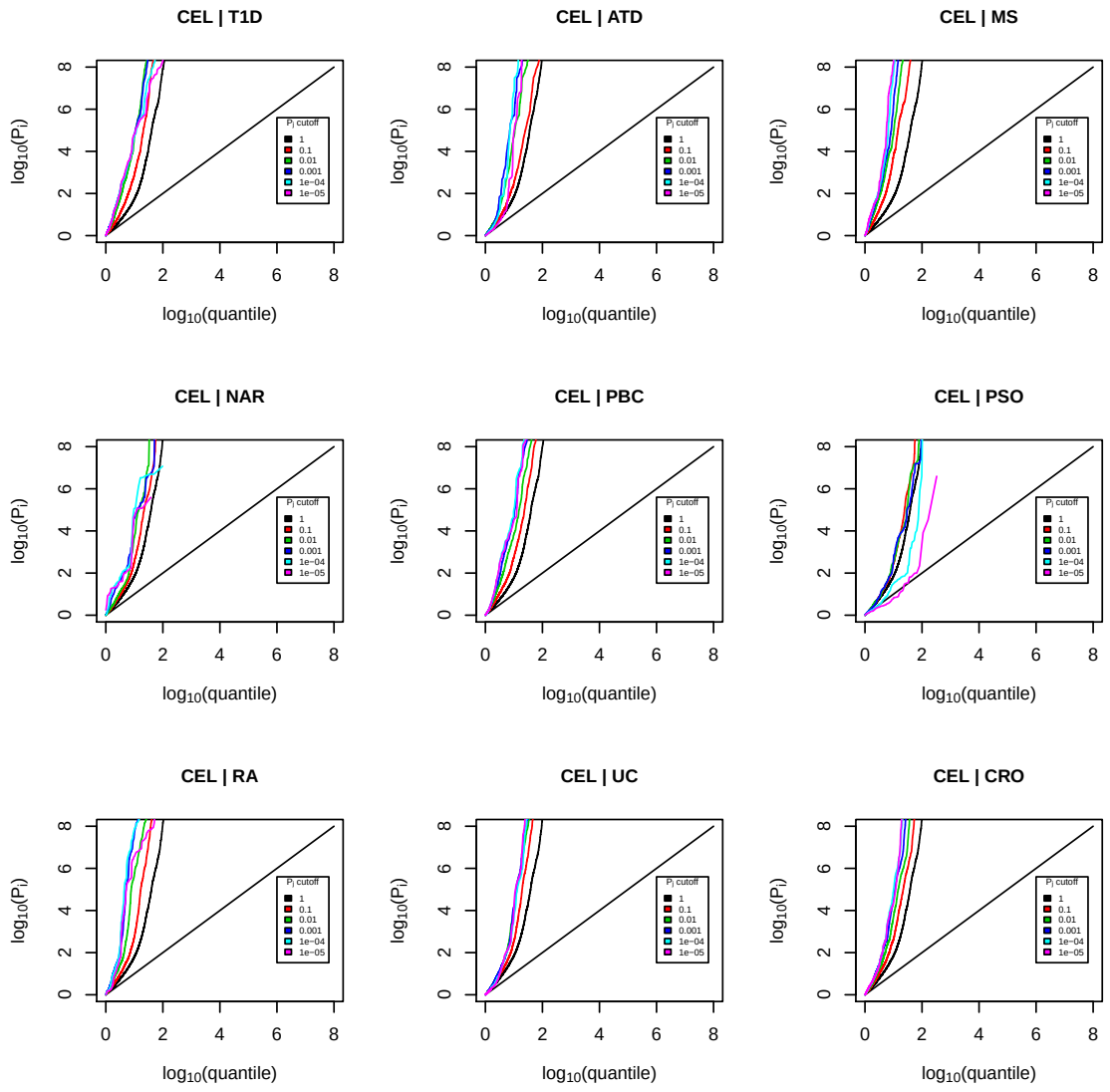


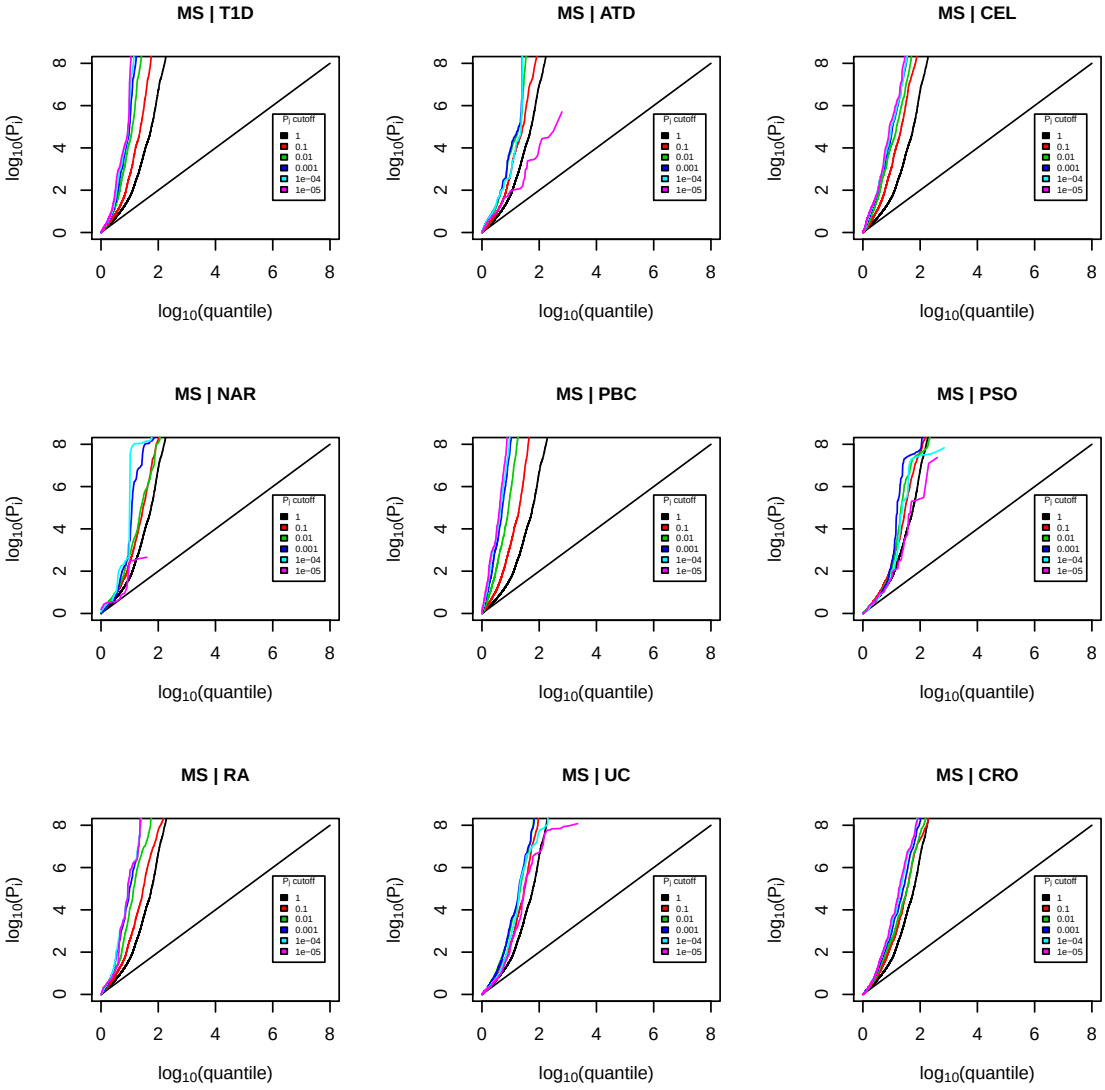
Fig. A.7 Distribution of $\log(M)$ amongst null SNPs. M is proportional to the variance of the log odds ratio from TDT data, defined as $\hat{\sigma}^2 f(1-f)$), where f is the minor allele frequency amongst null SNPs, and $\hat{\sigma}$ is the standard error. Equating the median of M with a known expression for variance of the log odds ratio in a case-control study enables back-calculation of the effective number of cases and controls. This technique was used for computing the number of cases and controls in the T1D study, for which p values were obtained from a meta-analysis of case-control and TDT data.

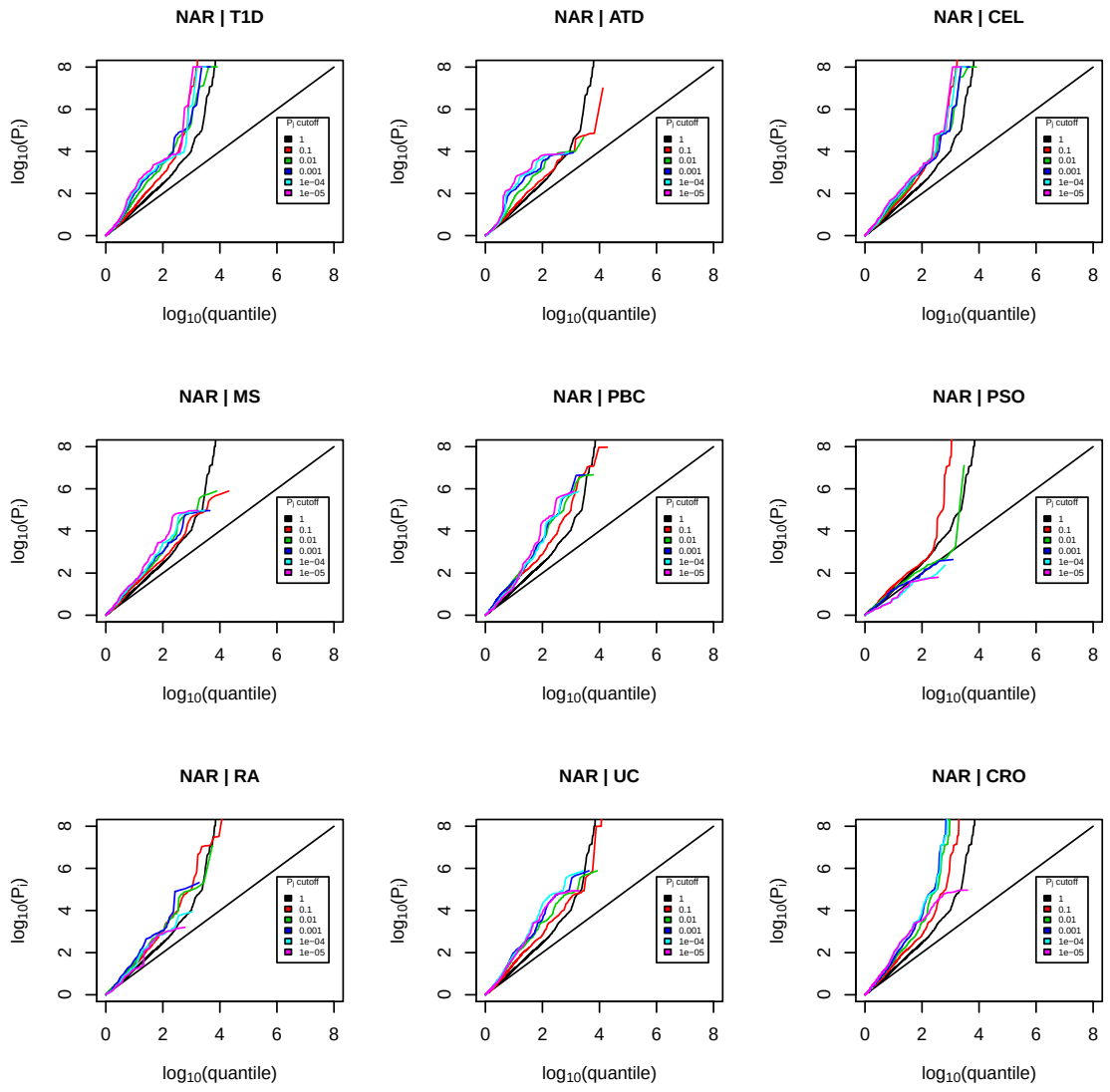
Fig. A.8 Q-Q plots labelled "$i|j$", where i is the principal phenotype and j the conditional phenotype. Y axes show $\log_{10}(p_i')$; X axes show log quantile (rank) of p values in various sets of SNPs. Each colour corresponds to the Q-Q plot for p_i amongst only SNPs such that p_j is less than a certain cutoff, with the black line corresponding to the Q-Q plot for all SNPs. P values for the principal phenotype are adjusted for the effect of shared controls between studies. A leftward shift with decreasing p_j cutoff indicates enrichment of SNP sets from conditioning on degrees of association with a conditional phenotype, probably due to pleiotropic effects between phenotypes. Because the studies used the ImmunoChip, which covers only potential autoimmune-associated regions, the black line also shows considerable enrichment compared to quantiles.

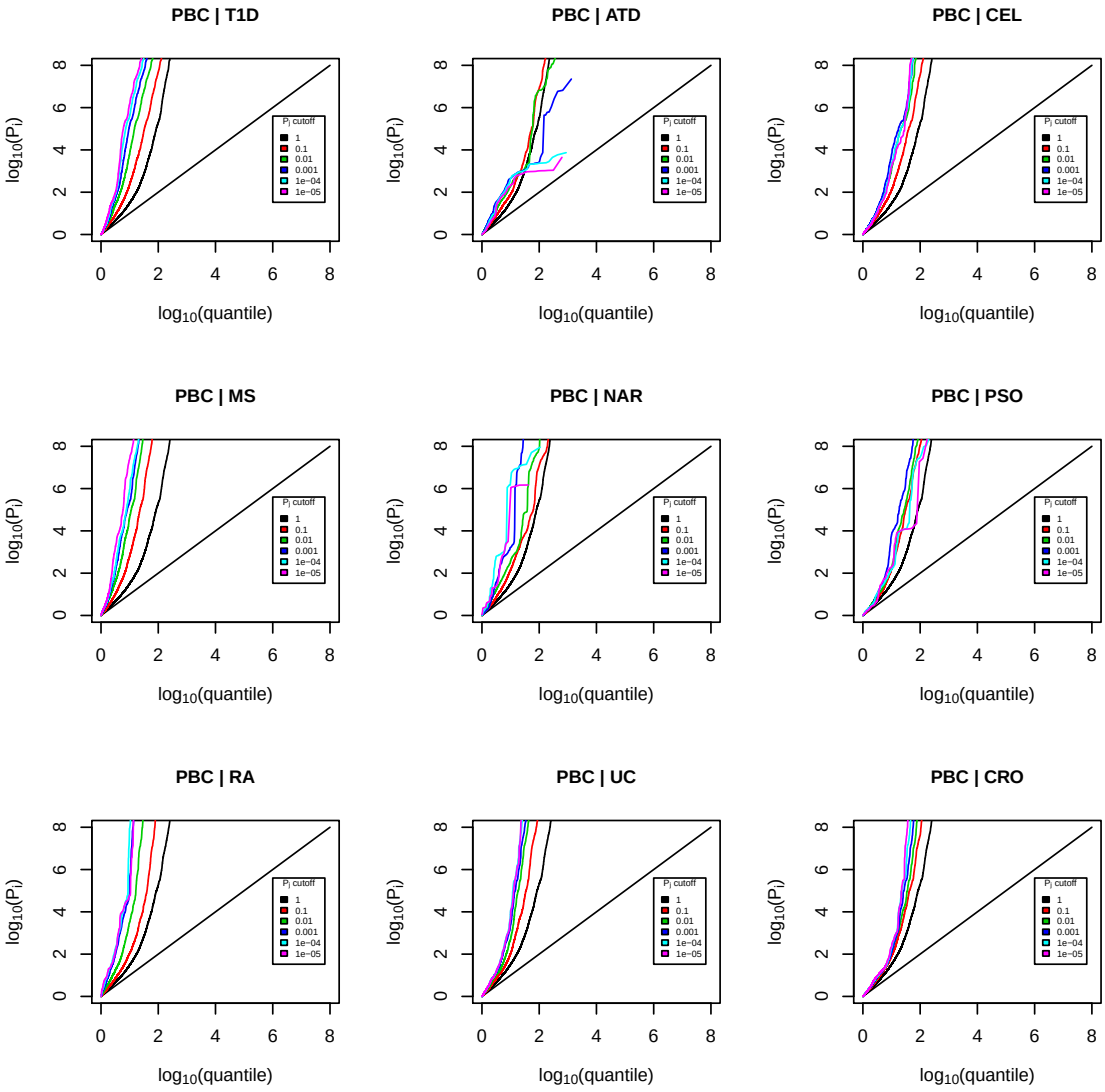


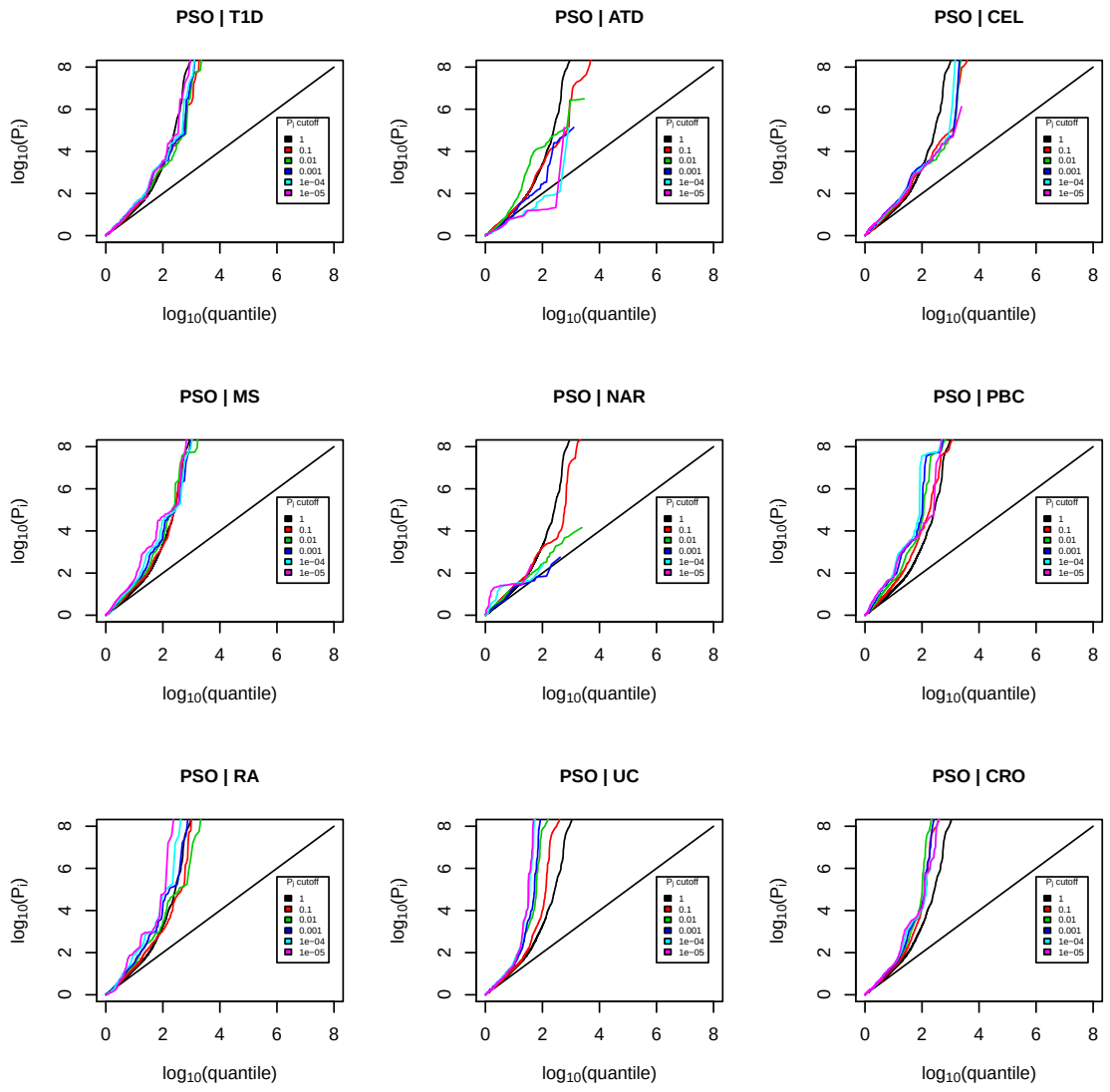

ATD | T1D
ATD | CEL
ATD | MS
ATD | NAR
ATD | PBC
ATD | PSO
ATD | RA
ATD | UC
ATD | CRO
log10(Pi)
log10(quantile)
Pj cutoff
1
0.1
0.01
0.001
1e−04
1e−05

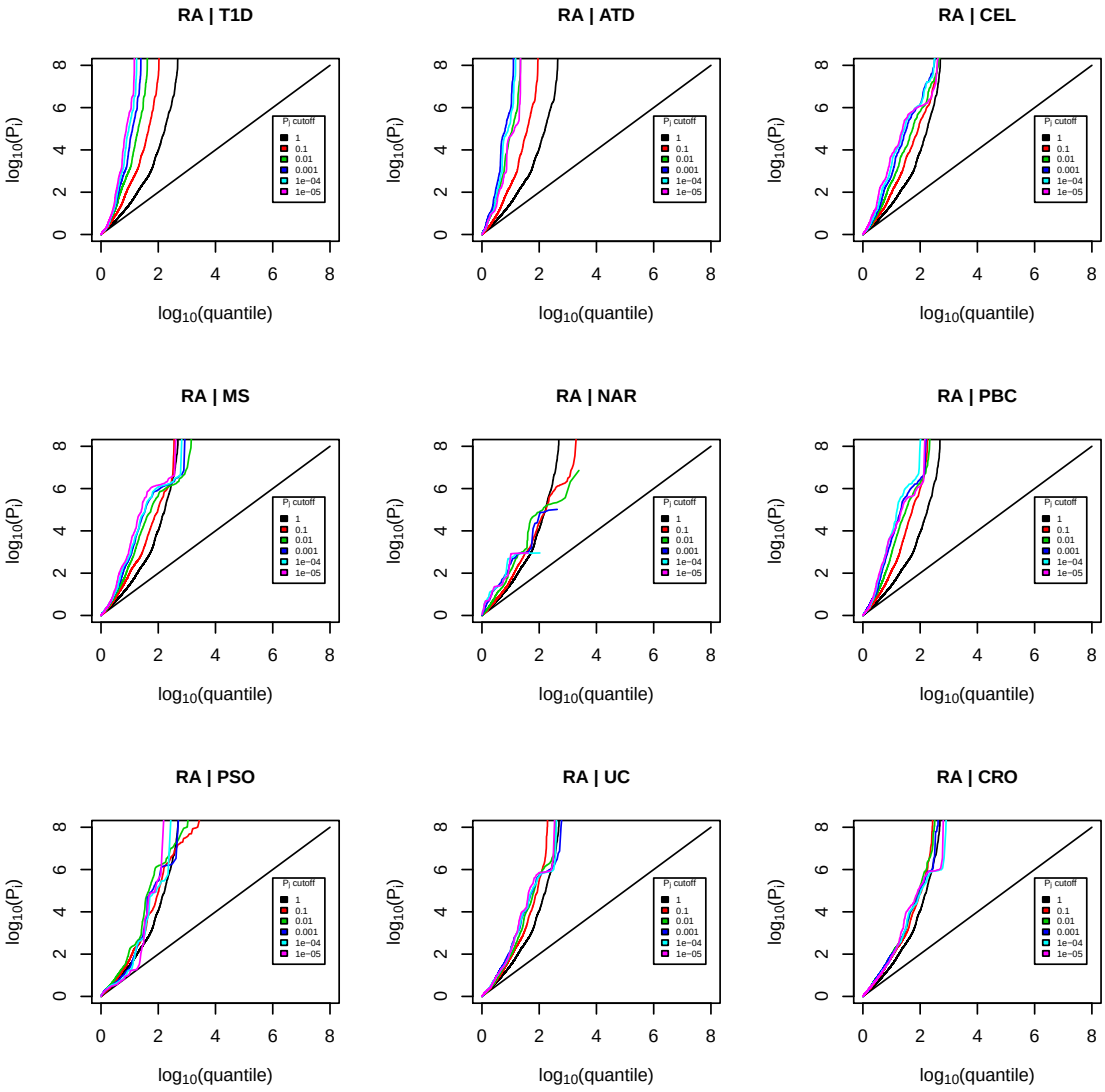

CEL | T1D
CEL | ATD
CEL | MS
CEL | NAR
CEL | PBC
CEL | PSO
CEL | RA
CEL | UC
CEL | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05

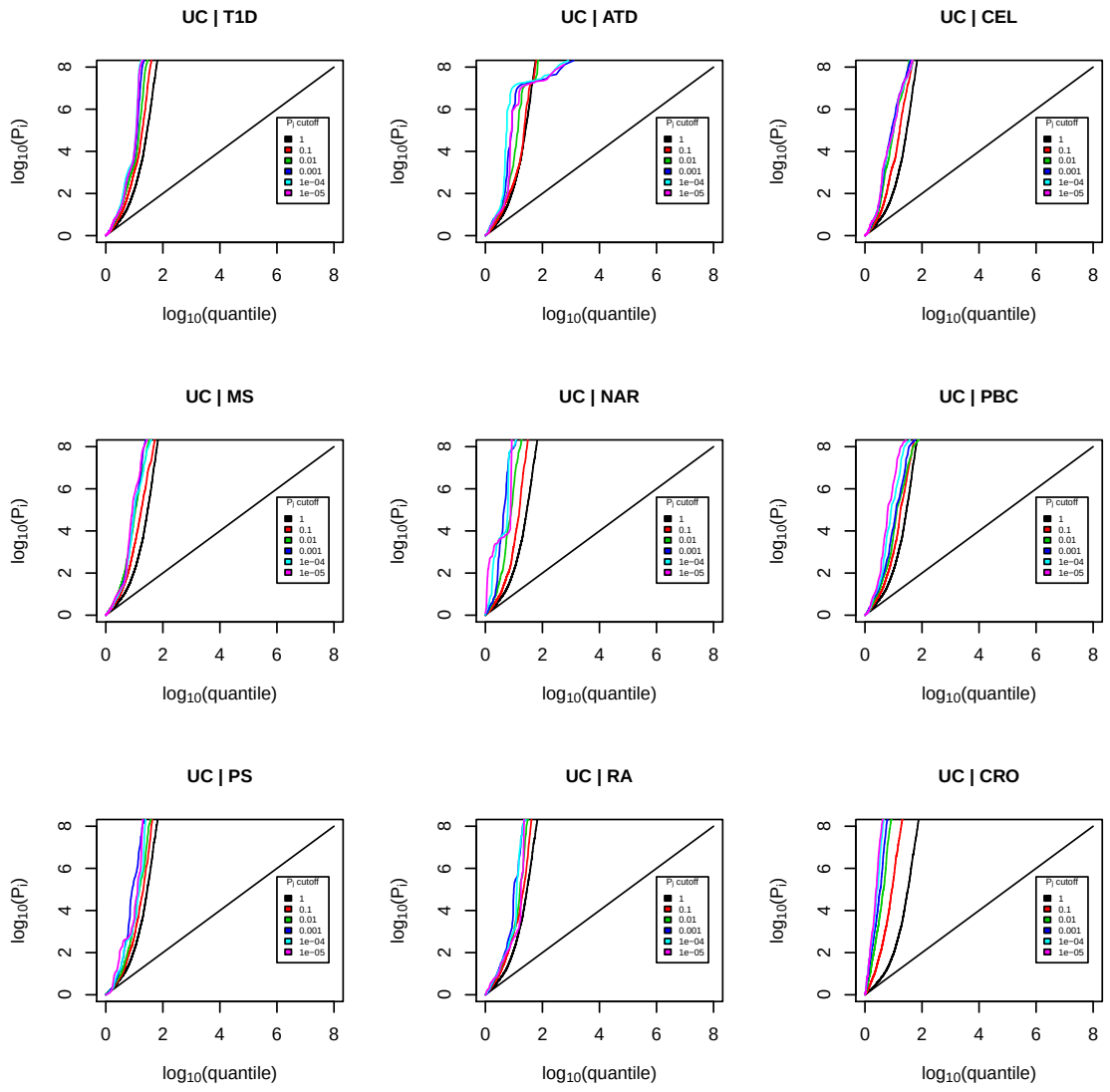

MS | T1D
MS | ATD
MS | CEL
MS | NAR
MS | PBC
MS | PSO
MS | RA
MS | UC
MS | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05
0
2
4
6
8

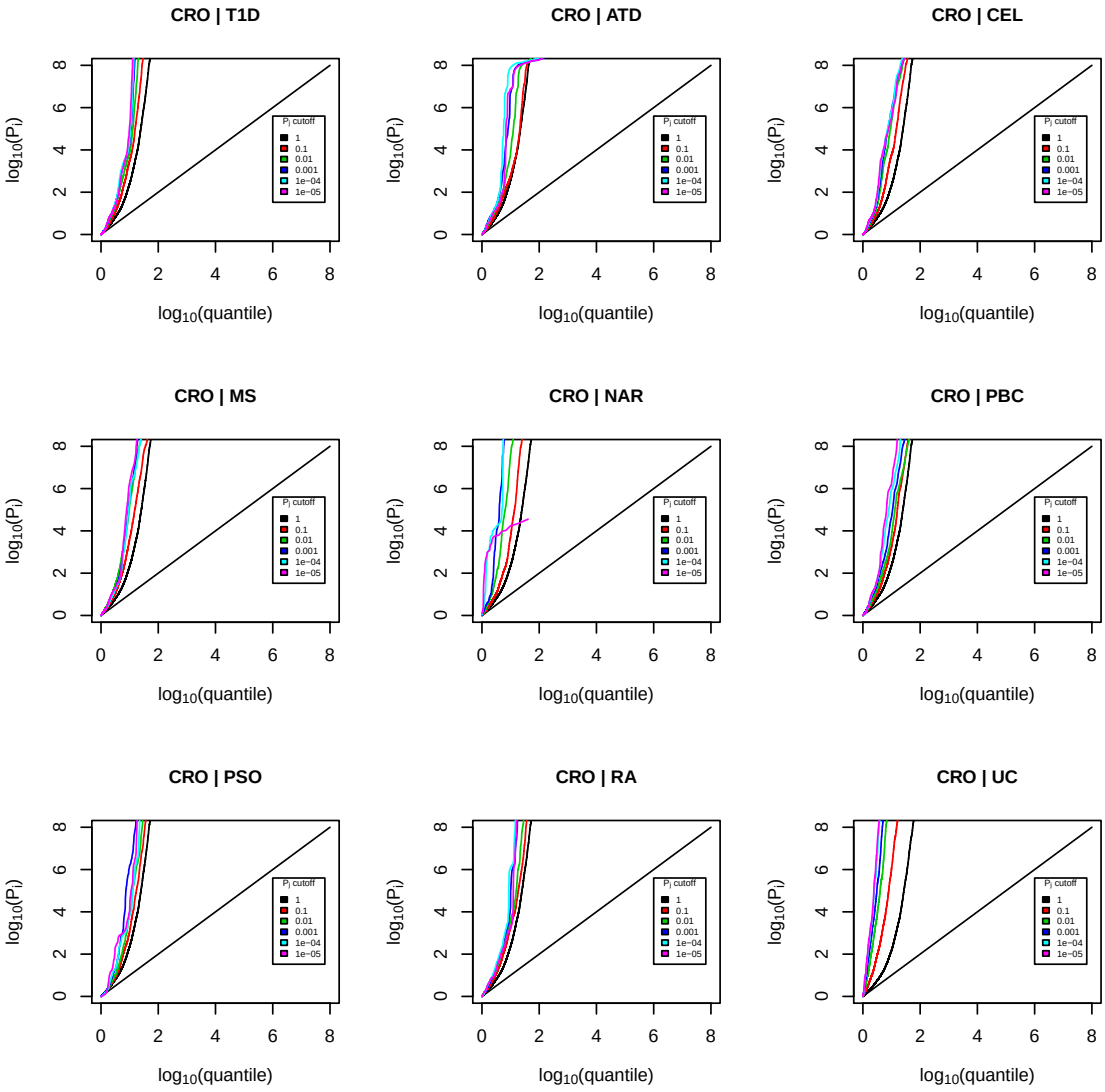

NAR | T1D
NAR | ATD
NAR | CEL
NAR | MS
NAR | PBC
NAR | PSO
NAR | RA
NAR | UC
NAR | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05
0
2
4
6
8


PBC | T1D
PBC | ATD
PBC | CEL
PBC | MS
PBC | NAR
PBC | PSO
PBC | RA
PBC | UC
PBC | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05


PSO | T1D
PSO | ATD
PSO | CEL
PSO | MS
PSO | NAR
PSO | PBC
PSO | RA
PSO | UC
PSO | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05
0
2
4
6
8


RA | T1D
RA | ATD
RA | CEL
RA | MS
RA | NAR
RA | PBC
RA | PSO
RA | UC
RA | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05
0
2
4
6
8


UC | T1D
UC | ATD
UC | CEL
UC | MS
UC | NAR
UC | PBC
UC | PS
UC | RA
UC | CRO
log10(Pi)
log10(quantile)
Pi cutoff
1
0.1
0.01
0.001
1e−04
1e−05


CRO | T1D
CRO | ATD
CRO | CEL
CRO | MS
CRO | NAR
CRO | PBC
CRO | PSO
CRO | RA
CRO | UC
log10(Pi)
log10(quantile)
Pj cutoff
1
0.1
0.01
0.001
1e−04
1e−05

Appendix B

Supplementary work for chapter 4

B.1 Supplementary note

B.1.1 Covariance between Z scores due to shared samples

The matching of type-1 error rates between methods relies on establishing the covariance between triples of z-scores under $H_0^{\cap}$. The covariance can be readily estimated when z-scores are assumed to be derived from tests of equality of binomial random variables m_i.

If strata or covariates are used, either an assumption must be made that computed z-scores are well-approximated by comparisons of binomial proportions, or covariances must be estimated allowing for the covariate or strata structure. Estimates can be made analytically in some circumstances, but can also be made empirically either by using known non-associated variants or by simulating variants with the same covariate structure.

The presence of strata or covariates also affects the values ζ_i, and if the effects of covariates are large, the approximations in equations 4.9 in the main paper may be poor. Values ζ_i can be estimated as functions of allelic differences by simulating variants with the same covariate structure.

No covariates or stratification

Assume study i and j have n_0^i, n_0^j controls and n_1^i, n_1^j cases respectively, of which n_0^{ij} controls and n_1^{ij} cases are shared between both studies. Let m_0, m_1, m_0', m_1', denote the observed allele frequencies of a SNP in the respective cohort, and $\mu_0, \mu_1, \mu_0', \mu_1'$ the expected allele frequency.

As for section 4.4 in chapter 4, define z_x for $x \in \{d, r, s, m, c\}$ as the signed z-score $(\pm\Phi^{-1}(p_x/2))$ corresponding to p_x, and z_x for $x \in \alpha, \beta, \beta^*, \gamma$ as the positive corresponding threshold $-\Phi^{-1}(x/2)$, where Φ, Φ^{-1} are the standard normal CDF and quantile functions.

If no strata or covariates are used in the calculation of summary statistics, z scores z_d, z_r, z_s, z_m are asymptotically proportional to the allelic differences $m_1 - m_0$, $m_1' - m_0'$, $m_1' - \frac{m_0' n_0' + m_0 n_0}{n_0 + n_0'}$, $\frac{m_1 n_1 + m_1' n_1'}{n_1 + n_1'} - \frac{m_0' n_0' + m_0 n_0}{n_0 + n_0'}$ respectively, since z scores are monotonic with allelic differences and allelic differences are asymptotically normal. Since m_0, m_1, m_0', m_1' are independent and asymptotically normal the multivariate random variables (z_d, z_r, z_m) and (z_d, z_s, z_m) have multivariate normal distributions.

For studies i on n_{0i}, n_{1i} controls and cases and j on n_{0j}, n_{1j} controls and cases in which n_{0ij} and n_{1ij} controls and cases are shared between studies, the correlation between the observed allelic differences $m_{1i} - m_{0i}$, $m_{1j} - m_{0j}$ is asymptotically given by

$$\mathrm{cor}(m_{1i} - m_{0i}, m_{1j} - m_{0j}) = \frac{n_{0i} n_{0j} n_{1ij} + n_{1i} n_{1j} n_{0ij}}{n_{0i} n_{0j} n_{1i} n_{1j} \sqrt{\frac{1}{n_{0i}} + \frac{1}{n_{0j}}} \sqrt{\frac{1}{n_{0j}} + \frac{1}{n_{1j}}}} \tag{B.1}$$

This holds under $H_0^{\cap}$ (no allelic differences between cohorts) and approximately holds in general. Expressions for ρ_{ds}, ρ_{dm}, ρ_{rm} and ρ_{sm} may be derived in terms of n_0, n_1, n_0', and n_1'. Specifically

$$\begin{aligned} \det(\Sigma_A) &= 1 - \rho_{dm}^2 - \rho_{rm}^2 \\ &= \frac{(n_0 n_1' - n_0' n_1)^2}{(n_0 + n_0')(n_1 + n_1')(n_0 + n_1)(n_0' + n_1')} \end{aligned} \tag{B.2}$$

$$\begin{aligned} \det(\Sigma_B) &= 1 - \rho_{dm}^2 - \rho_{ds}^2 - \rho_{sm}^2 + 2\rho_{dm}\rho_{ds}\rho_{sm} \\ &= \frac{n_0' n_1^2}{(n_0 + n_1)(n_0 + n_0' + n_1')(n_1 + n_1')} \end{aligned} \tag{B.3}$$

so Σ_A is singular if $\frac{n_0}{n_1} = \frac{n_0'}{n_1'}$, and Σ_B if $n_0' n_1 = 0$.

Z scores with stratification

If computation of Z scores is performed with correction for strata or covariates, formula B.1 will not asymptotically hold under $H_0^{\cap}$ and may be a poor approximation to the true covariance. The true covariance can be computed in some cases.

If samples are divided into strata $1, 2, \ldots s$, and n_{pi}^r, m_{pi}^r, μ_{pi}^r denote the number of samples and observed and expected minor allele frequencies in cohort p, study i, stratum r respectively, then the z score z_i for study i is asymptotically given by

$$z_i = \sum_{r=1}^{s} \alpha_{ir} \left(m_{1i}^r - m_{0i}^r\right) \tag{B.4}$$

for positive values α_{ir} depending on the values n^r_{pi}. If the Cochran-Mantel-Hanszel test is used, then

$$\alpha_{ir} \propto \frac{n^r_{0i} n^r_{1i}}{n^r_{0i} + n^r_{1i}} \tag{B.5}$$

Suppose that n^r_{0ij} controls and n^r_{1ij} cases are shared between studies i and j in stratum r. Since the values m^r_{pi} are dependent only within the same values of p and r, the correlation between z_i and z_j under the null hypothesis $\mu^r_{0i} \equiv \mu^r_{1i}, \mu^r_{0i} \equiv \mu^r_{1i}$ is given by

$$\begin{aligned} \mathrm{cor}(z_i, z_j) &= \frac{\sum \alpha_{ir}\alpha_{jr} \mathrm{cov}(m^r_{1i} - m^r_{0i}, m^r_{1j} - m^r_{0j})}{\sqrt{\mathrm{var}(m^r_{1i} - m^r_{0i})\mathrm{var}(m^r_{1j} - m^r_{0j})}} \\ &\approx \frac{\sum \alpha_{ir}\alpha_{jr} \left(\frac{n^r_{0ij}}{n^r_{0i} n^r_{0j}} + \frac{n^r_{1ij}}{n^r_{1i} n^r_{1j}} \right)}{\sqrt{\left(\sum \alpha^2_{ir} \left(\frac{1}{n^r_{0i}} + \frac{1}{n^r_{1i}} \right) \right) \left(\sum \alpha^2_{jr} \left(\frac{1}{n^r_{0j}} + \frac{1}{n^r_{1j}} \right) \right)}} \end{aligned} \tag{B.6}$$

where all sums are implicitly only over the values of $r \in 1..s$ for which the relevant values of n^r_{pi} are positive.

Z scores with covariates

If z scores are computed adjusting for one or more covariates, the estimation of correlation is more difficult. Assume that in a case cohort C_1 and a control cohort C_0 the values of some covariate(s) x have different known distributions f_1, f_0, and that numerical genotypes g ($g \in \{0,1,2\}$) at some SNP of interest may vary with x. Assume the cohorts are large and that f_1, f_0, and $E(g|x)$ are continuous functions of x.

Let g^k_p denote the genotype of individual k in cohort p ($p \in C_0, C_1$) and x^k_p denote covariate value(s), where g^k_p is an observation of a random variable g. An idealised z-score testing association of g with case/control status can be considered a function of the values g^k_p, x^k_p which should be monotonic with each g^k_p and have expectation under $H^\cap_0$ if x is independent of case/control status, whatever the form of the function $E(g|x)$. Because individual genotypes are assumed to be independent between individuals, cross-terms of the form $\prod_i g_i$ should carry no additional information from singleton genotypes, so I assume that z can be decomposed into a weighted linear sum of individual genotypes:

$$z \propto \frac{1}{|C_1|} \sum_{k \in C_1} c^k_1 g^k_1 - \frac{1}{|C_0|} \sum_{k \in C_0} c^k_0 g^k_0 \tag{B.7}$$

where the (positive) values c_1^k, c_0^k depend only on the values $x_1^{\cdot}, x_0^{\cdot}$; that is, not on the relationship between g and x, and the constant of proportionality depends on only on the observed allele frequency. Let function $c_0(x), c_1(x)$ denote the values of c_i corresponding to covariate value(s) x in C_0, C_1.

For a null SNP, $E(g|x)$ is independent of case/control status, but may take any (continuous) form. We have

$$\begin{aligned} E(z) &\propto E\left(\frac{1}{|C_1|}\sum_{i\in C_1} c_i g_i - \frac{1}{|C_0|}\sum_{i\in C_0} c_i g_i\right) \\ \lim_{|C_0|,|C_1|\to\infty} E(z) &\propto \int c_1(x) f_1(x) E(g|x) dx - \int c_0(x) f_0(x) E(g|x) \\ &\propto \int \left(c_1(x) f_1(x) - c_0(x) f_0(x)\right) E(g|x) dx \end{aligned} \tag{B.8}$$

From a standard result from the calculus of variations, $E(z) = 0$ implies that

$$c_1(x) f_1(x) - c_0(x) f_0(x) \equiv 0 \implies c_1(x) \propto \frac{f(x)}{f_1(x)}, c_0(x) \propto \frac{f(x)}{f_0(x)} \tag{B.9}$$

for some function f, so the values c_1^k, c_0^k must effectively re-weight the contribution of individuals to a common density $f(x)$ across x. The procedure of weighting observation k in this way is analogous to a limiting case of stratification, in which weights are defined by the frequency of stratum r (see discussion of strata above). For a constant allelic difference across the range of x, the best common distribution to 'map to' does not depend on the relationship between g and x, and hence the best values of c_i should be constant for all functions $E(g|x)$.

Let z_q denote a z-score for study q; n_{pq}, $f_{pq} = f_{pq}(x)$ and C_{pq} denote the number of samples, density function of x, and set of samples in cohort p, study q; g_{pq}^i and c_{pq}^k denote the normalised genotype of sample k in cohort p, study q and its coefficient in z_q; n_{0s}, n_{1s}, f_{0s}, f_{1s} and C_0^s, C_1^s the number of shared controls/cases between studies, the density of x amongst the shared samples, and the sets of shared samples; and f_q the common density function to

which cases and controls are weighted in study q (equation B.9). Then

$$\mathrm{cov}(z_i, z_j) \approx \frac{\frac{1}{n_{0i}n_{0j}}\sum_{k\in C_0^s} c_{0i}^k c_{0j}^k + \frac{1}{n_{1i}n_{1j}}\sum_{k\in C_1^s} c_{1i}^k c_{1j}^k}{\sqrt{\frac{1}{n_{1i}^2}\sum_{k\in C_{1i}}(c_{1i}^k)^2 + \frac{1}{n_{0i}^2}\sum_{k\in C_{0i}}(c_{0i}^k)^2}\sqrt{\frac{1}{n_{1j}^2}\sum_{k\in C_{1j}}(c_{1j}^k)^2 + \frac{1}{n_{0j}^2}\sum_{k\in C_{0j}}(c_{0j}^k)^2}} \tag{B.10}$$

$$\to \frac{\frac{n_{0s}}{n_{0i}n_{0j}}\int f_{0s}(x)\frac{f_i(x)f_j(x)}{f_{0i}(x)f_{0j}(x)}dx + \frac{n_{1s}}{n_{1i}n_{1j}}\int f_{1s}(x)\frac{f_i(x)f_j(x)}{f_{1i}(x)f_{1j}(x)dx}}{\sqrt{\frac{1}{n_{0i}}\int\frac{f_i(x)^2}{f_{0i}(x)}dx + \frac{1}{n_{1i}}\int\frac{f_i(x)^2}{f_{1i}(x)}dx}\sqrt{\frac{1}{n_{0j}}\int\frac{f_j(x)^2}{f_{0j}(x)}dx + \frac{1}{n_{1j}}\int\frac{f_j(x)^2}{f_{1j}(x)}dx}} \tag{B.11}$$

with integrals over the domain of x, and the limit as sample sizes tend to infinity while ratios between them remain constant.

Logistic regression models with continuous covariates can only model simple (generally linear) relationships between c_i and x_i, and property B.9 may not hold. If the values c_{pq}^k are known, the correlation can be determined using equation B.10. If not, some methods for estimating correlation are outlined below.

Practical estimation of covariance

Although the asymptotic correlation between z scores may be intractable, as long as the z score permits an expansion of the form B.7, the correlation will be nearly invariant with allele frequency and change only minimally for SNPs associated with the covariate.

In practical terms, one method to estimate the correlation between z scores is to simply use the sample correlation at a set of variants presumed to be not associated with the main trait of interest. This approach may be unreliable and have limited accuracy due to the difficulty of identifying such variants Another option is to permute existing genotypes without permuting covariates, and compute correlation between resultant z scores. This has the disadvantage that it is difficult to permute whilst maintaining potential relationships between genotypes and confounders.

Since the correlation should only depend on the sample sizes and structure of covariate distributions, a more convenient and powerful method is to simply simulate sets of genotypes un-associated with the trait, but potentially associated with covariates in a range of different ways, and compute correlation between the resultant z scores. Given the shortcomings of standard methods such as logistic regression in fully accounting for covariate effects, this is an advisable procedure in any analysis including covariates.

All results in the main paper which require conditions on sample sizes are only approximate when using studies with stratification or covariates, with the approximation worsening with greater differences in covariate values between groups and lower effective sample sizes.

B.1.2 Properties of β^*

Asymptotic properties of β^*

In this appendix, an asymptotic approximation is established for $\beta^*$and it is shown that $\beta^* > \beta$ for all n_0^i, n_0^j, n_1^i, n_1^j, z_α, z_γ. Define Σ_A and Σ_B as per equations 4.1 in the main paper, and note that Σ_A and Σ_B only differ in their middle row/column. Further define

$$\Sigma_{dm} = \text{var}\left((z_d\, z_m)^t | H_0^{\cup}\right) = \begin{pmatrix} 1 & \rho_{dm} \\ \rho_{dm} & 1 \end{pmatrix} \tag{B.12}$$

Let $(z'_\alpha\, z'_\gamma)$ be the point in $\{z_d > z_\alpha, z_m > z_\gamma\}$ at minimal Mahalanobis distance from the origin with respect to Σ_{dm} (ie, minimal $(z_d\, z_m)\Sigma_{dm}^{-1}(z_d\, z_m)^t$). Then for $z'_\gamma - \rho_{dm} z'_\alpha$ held constant, we have

$$\lim_{z'_\gamma \to \infty / z'_\alpha \to \infty} \frac{\sqrt{|\Sigma_A|}\left((\rho_{ds}\rho_{dm} - \rho_{sm})z'_\gamma + (\rho_{dm}\rho_{sm} - \rho_{ds})z'_\alpha + |\Sigma_{dm}| z_{\beta^*}\right)}{\sqrt{|\Sigma_B|}\left(-\rho_{rm} z'_\gamma + \rho_{dm}\rho_{rm} z'_\alpha + |\Sigma_{dm}| z_\beta\right)} = 1 \tag{B.13}$$

Specifically, for β^* defined as per equation 4.3, we have

$$\lim_{\alpha \to 0} \frac{z_{\beta^*}}{\sqrt{1-\rho_{ds}^2} z_\beta + \rho_{ds} z_\alpha} = 1 \tag{B.14}$$

and $z_{\beta^*} > \max(\beta, \sqrt{1-\rho_{ds}^2} z_\beta + \rho z_\alpha)$ for all z_α. To show this, I firstly establish the following lemma and corollary:

Lemma 1. *Let Σ be a positive definite matrix of dimension N, $\mathbf{x}$ be the vector $(x_1\, x_2 ... x_n)^t$, $\mathbf{A_1}$, $\mathbf{A_0}$, and $\mathbf{Z} = (z_1\, z_2 ... z_n)^t$ constant vectors of dimension N with $\mathbf{A_1} \neq \mathbf{A_0} \neq 0$, C_0 a constant, and R the (closed) region $x_1 \geq z_1, x_2 \geq z_2, , , x_N \geq z_N$.*

Define C as the (unique) value satisfying

$$\int_R e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}} \left(\Phi(\mathbf{A_1^t x} + C) - \Phi(\mathbf{A_0^t x} + C_0)\right) dx_1 dx_2 ... dx_N = 0 \tag{B.15}$$

Denote $\mathbf{y} = (y_1\, y_2 ... y_N)$ *as the point in R at minimal Mahalanobis distance* $M(\mathbf{y})$ *from the origin with respect to* Σ *(usually,* $\mathbf{y} = \mathbf{Z}$*). Consider all regions R for which the corresponding value of* $\mathbf{y}$ *lies on the hyperplane* $\mathbf{A_0^t y} + C_0' = 0$*,* $C_0' \neq C_0$*. Then*

$$\lim_{M(y)\to\infty | \mathbf{A_0^t y}+C_0'=0} \frac{\mathbf{A_1^t y}+C}{\mathbf{A_0^t y}+C_0} = \lim_{M(y)\to\infty | \mathbf{A_0^t y}+C_0'=0} \frac{\mathbf{A_1^t y}+C}{C_0 - C_0'} = 1 \tag{B.16}$$

Proof. The value C is unique since the function $\Phi(\mathbf{A_1^t x} + C)$ is continuous and monotonically increasing in C for all $\mathbf{x}$, and hence so is the integrand (and integral).

I proceed from the formal definition of a limit

$$\forall \varepsilon > 0\, \exists Y \,|\, \left(M(y) > Y \implies \left| \frac{\mathbf{A_1^t y}+C}{\mathbf{A_0^t y}+C_0} - 1 \right| < \varepsilon \right) \tag{B.17}$$

Because $\mathbf{A_0^t y} + C_0' = 0$, the right-hand side is equivalent to

$$(1-\varepsilon)(C_0 - C_0') - \mathbf{A_1^t y} \leq C \leq (1+\varepsilon)(C_0 - C_0') - \mathbf{A_1^t y} \tag{B.18}$$

I will show that there exists Y such that $M(y) > Y$ implies that when C takes values at the endpoints of the interval in the integral B.18, the integral B.15 takes different signs. Since the integral is increasing in C and must be 0, C must lie in the interval in B.18 for $M(y) > Y$.

If C takes the upper value, then at $\mathbf{x} = \mathbf{y}$, the value of the integrand is

$$e^{-\frac{1}{2}M(\mathbf{y})} \left(\Phi((1+\varepsilon)(C_0 - C_0')) - \Phi(C_0 - C_0') \right) \tag{B.19}$$

the sign of which depends on the sign of $C_0 - C_0'$. I shall assume it is positive (with analogous arguments if it is negative). Because $\varepsilon > 0$, point $\mathbf{y}$ does not lie on the hyperplane $(\mathbf{A_1^t} - \mathbf{A_0^t})\mathbf{x} + (1+\varepsilon)(C_0 - C_0') - \mathbf{A_1^t y} - C_0 = 0$ (on which the integrand of B.15 is 0). The distance from $\mathbf{y}$ to the hyperplane is given by

$$\begin{aligned} D &= \frac{|(\mathbf{A_1^t} - \mathbf{A_0^t})\mathbf{y} + (1+\varepsilon)(C_0 - C_0') - \mathbf{A_1^t y} - C_0|}{||\mathbf{A_1^t} - \mathbf{A_0^t}||} \\ &= \frac{|(1+\varepsilon)(C_0 - C_0') - C_0 - C_0'|}{||\mathbf{A_1^t} - \mathbf{A_0^t}||} \end{aligned} \tag{B.20}$$

which is independent of $\mathbf{y}$. Consider a hypersphere centred at $\mathbf{y}$ of radius $d < D$. Each point in the hypersphere can be expressed as $\mathbf{y} + \kappa$ with $|\kappa| \leq d$, so within the hypersphere we

have

$$\begin{aligned}\Phi(\mathbf{A_1^t}\mathbf{x}+C)-\Phi(\mathbf{A_0^t}\mathbf{x}+C_0) &= \Phi(\mathbf{A_1^t}(\mathbf{y}+\kappa)+(1+\varepsilon)(C_0-C_0')-\mathbf{A_1^t}\mathbf{y}) \\ &\quad -\Phi(\mathbf{A_0^t}(\mathbf{y}+\kappa)+C_0) \\ &= \Phi\left((1+\varepsilon)(C_0-C_0')+\mathbf{A_1^t}\kappa\right) \\ &\quad +\Phi\left((C_0-C_0')+\mathbf{A_0^t}\kappa\right) \\ &\geq \Phi\left((1+\varepsilon)(C_0-C_0')+|\mathbf{A_1^t}|d\right) \\ &\quad +\Phi\left((C_0-C_0')-|\mathbf{A_0^t}|d\right) \end{aligned} \tag{B.21}$$

Thus d can be chosen independently of $\mathbf{y}$ such that $\Phi(\mathbf{A_1^t}\mathbf{x}+C)-\Phi(\mathbf{A_0^t}\mathbf{x}+C_0)$ is bounded below in the hypersphere by a constant X also independent of $\mathbf{y}$. The function $\Phi(\mathbf{A_1^t}\mathbf{x}+C)-\Phi(\mathbf{A_0^t}\mathbf{x}+C_0)$ is obviously bounded by ± 2. Let R' be the intersection of R and the hypersphere. The integral B.15 now satisfies

$$\begin{aligned}&\int_R e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}\left(\Phi(\mathbf{A_1^t}\mathbf{x}+C)-\Phi(\mathbf{A_0^t}\mathbf{x}+C_0)\right)dx_1dx_2...dx_N \\ &= \int_{R'} e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}\left(\Phi(\mathbf{A_1^t}\mathbf{x}+C)-\Phi(\mathbf{A_0^t}\mathbf{x}+C_0)\right)dx_1dx_2...dx_N \\ &\quad + \int_{R\setminus R'} e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}\left(\Phi(\mathbf{A_1^t}\mathbf{x}+C)-\Phi(\mathbf{A_0^t}\mathbf{x}+C_0)\right)dx_1dx_2...dx_N \\ &> X\int_{R'} e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}dx_1dx_2...dx_N \\ &\quad -2\int_{R\setminus R'} e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}dx_1dx_2...dx_N \end{aligned} \tag{B.22}$$

Because d (the radius of the hypersphere) does not depend on $\mathbf{y}$, by the properties of the Gaussian integral a value M_+ can be chosen such that $M(y) > M_+$ implies that the ratio

$$\frac{\int_{R'} e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}dx_1dx_2...dx_N}{\int_{R\setminus R'} e^{-\frac{1}{2}\mathbf{x^t}\Sigma^{-1}\mathbf{x}}dx_1dx_2...dx_N} \tag{B.23}$$

is arbitrarily large (namely, $> 2/X$), and hence integral B.22 is positive. In a similar way, a value M_- can be chosen such that if C takes the lower value of interval B.18, the integral is negative for $M(y) > M_-$. For $M(y) > \max(M_+, M_-)$, the value of C satisfying equation B.15 lies within the interval B.18, and the limit is established.

□

Corollary 3. *Given $b,c,y \in \mathbb{R}^+$, define a such that*

$$\int_y^\infty e^{-\frac{x^2}{2}} \left(\Phi(c) - \Phi(a-bx)\right) dx = 0 \tag{B.24}$$

then

$$\lim_{y\to\infty} \frac{a}{by+c} = 1 \tag{B.25}$$

and $a > by + c \; \forall \, y$

Proof. Note firstly that the function $\Phi(c) - \Phi(a-bx)$ is increasing for all x. If the integral is 0, the (smooth) integrand must cross 0 at some finite $x \in (y,\infty)$, and hence its value at $x = y$ must be negative. As Φ is increasing, we have $\Phi(a-by) > \Phi(c) \implies a > by + c$

The proof of the limit proceeds in a similar way to the proof of the lemma above.

□

Now (recalling definition 4.2 in the main paper)

$$\begin{aligned}
&\int_{z_\alpha}^\infty \int_{z_\gamma}^\infty \int_{z_{\beta^*}}^\infty N_{\Sigma_B}\left((z_d \, z_s \, z_m)^t\right) dz_s dz_m dz_d \\
&= \int_{z_\alpha}^\infty \int_{z_\gamma}^\infty \int_{z_\beta}^\infty N_{\Sigma_A}\left((z_d \, z_r \, z_m)^t\right) dz_r dz_m dz_d \\
&\implies \int_{z_\alpha}^\infty \int_{z_\gamma}^\infty N_{\Sigma_{dm}}\left((z_d \, z_m)^t\right) \left(\Phi\left(a_1 z_d + b_1 z_m + c_1\right)\right. \\
&\qquad\qquad \left. - \Phi\left(a_0 z_d + b_0 z_m + c_0\right)\right) dz_d dz_m = 0
\end{aligned} \tag{B.26}$$

where

$$
\begin{aligned}
a_0 &= -\frac{\rho_{dm}\rho_{rm}}{\sqrt{|\Sigma_{dm}||\Sigma_A|}} \\
b_0 &= \frac{\rho_{rm}}{\sqrt{|\Sigma_{dm}||\Sigma_A|}} \\
c_0 &= -\sqrt{\frac{|\Sigma_{dm}|}{|\Sigma_A|}} z_\beta \\
a_1 &= \frac{\rho_{ds}-\rho_{dm}\rho_{sm}}{\sqrt{|\Sigma_{dm}||\Sigma_B|}} \\
b_1 &= \frac{\rho_{sm}-\rho_{ds}\rho_{dm}}{\sqrt{|\Sigma_{dm}||\Sigma_B|}} \\
c_1 &= -\sqrt{\frac{|\Sigma_{dm}|}{|\Sigma_B|}} z_{\beta^*}
\end{aligned} \tag{B.27}
$$

The asymptotic property of β^* follows from corollary 3.

If $\gamma = 1$, we have from definition 4.3 in the main paper

$$
\begin{aligned}
&\int_{z_\alpha}^{\infty}\int_{z_{\beta^*}}^{\infty} \frac{1}{2\pi\sqrt{1-\rho_{ds}^2}} \exp\left(-\frac{1}{2(1-\rho_{ds}^2)}\left(x^2+y^2-2\rho xy\right)\right) dxdy \\
&= \int_{z_\alpha}^{\infty}\int_{z_\beta}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(x^2+y^2\right)\right) dxdy \\
&\implies \int_{z_\alpha}^{\infty} e^{-\frac{y^2}{2}} \Phi\left(\frac{z_{\beta^*}-\rho_{ds}y}{\sqrt{1-\rho_{ds}^2}}\right) dy = \int_{z_\alpha}^{\infty} e^{-\frac{y^2}{2}} \Phi(z_\beta)
\end{aligned} \tag{B.28}
$$

from which the result follows from an application of lemma 1.

Size of β, β^* and $\beta^\perp$

To show that $\beta^* < \beta$, I show that if we replace z_β with z_{β^*} in the definition of c_0 in equation B.27, then the integral B.26 is positive. Since the value of the integral is decreasing with z_β^* (as $\Phi(x)$ is monotonically increasing in x) we must have $z_\beta^* > z_\beta$ if the integral is to be 0. A similar argument can be used to show that $\beta^\perp < \beta^*$. Denote by $I(z_d, z_m)$ the value of the integrand of B.26 with $z_{\beta^*} = z_\beta$.

Consider the line $a_1 z_d + b_1 z_m + c_1 = a_0 z_d + b_0 z_m + c_0$ on the (z_d, z_m) plane on which the integrand of B.26 is 0. The gradient of this line is

$$\frac{a_0 - a_1}{b_0 - b_1} = \frac{\sqrt{n_0^i (n_0 + n_0')n_1(n_0 + n_1)}}{\sqrt{(n_1 + n_1')(n_0 + n_0' + n_1 + n_1')}} \times \tag{B.29}$$

$$\frac{n_0' n_1 (n_0 + n_0' + n_1 + n_1') - (n_0 + n_0')|n_0' n_1 - n_0 n_1'|}{n_0'(n_0 + n_0')n_1(n_0 + n_1) - (n_0^2 + n_0 n_0' + n_0' n_1)|n_0' n_1 - n_0 n_1'|} \tag{B.30}$$

Since $|n_0' n_1 - n_0 n_1'| \geq (n_0' n_1 - n_0 n_1')$ the numerator of the second fraction is greater than or equal to

$$\begin{aligned} n_0' n_1 (n_0 + n_0' + n_1 + n_1') - (n_0 + n_0')(n_0' n_1 - n_0 n_1') &= n_0^2 n_1' + n_0'(n_1^2 + n_0 n_1' + n_1 n_1') \\ &> 0 \end{aligned} \tag{B.31}$$

and similarly the denominator is greater than or equal to

$$n_0(n_0^2 n_1' + n_0'(n_1^2 + n_0 n_1' + n_1 n_1')) > 0 \tag{B.32}$$

so the gradient is positive. If $b_1 - b_0 > 0$, $I(z_d, z_m)$ is positive if (z_d, z_m) falls above the line, and negative if below it; if $b_1 - b_0 < 0$, the other way around. Assume for the moment that $b_1 - b_0 < 0$.

If the point (z_α, z_γ) lies above the line, then since $I(z_d, z_m)$ is negative in the region $(-\infty, z_\alpha) \times (z_\gamma, \infty)$, we have

$$\begin{aligned} \int_{z_\alpha}^{\infty} \int_{z_\gamma}^{\infty} I(z_d, z_m) dz_d dz_m &\geq \int_{z_\alpha}^{\infty} \int_{z_\gamma}^{\infty} I(z_d, z_m) dz_d dz_m \\ &\quad + \int_{-\infty}^{z_\alpha} \int_{z_\gamma}^{\infty} I(z_d, z_m) dz_d dz_m \\ &= \int_{\infty}^{\infty} \int_{z_\gamma}^{\infty} I(z_d, z_m) dz_d dz_m \end{aligned} \tag{B.33}$$

If the point lies below the line, let $z_\gamma' > z_\gamma$ be defined such that the point (z_α, z_γ') lies on the line. Since $I(z_d, z_m)$ is positive in the region $(z_\alpha, \infty) \times (z_\gamma, z_\gamma')$ and negative in the region

$(-\infty, z_\alpha) \times (z_\gamma, \infty)$, we have

$$\begin{aligned}\int_{z_\alpha}^{\infty}\int_{z_\gamma}^{\infty} I(z_d, z_m)dz_d dz_m \geq & \int_{z_\alpha}^{\infty}\int_{z_\gamma}^{\infty} I(z_d, z_m)dz_d dz_m \\ & - \int_{z_\alpha}^{\infty}\int_{z_\gamma}^{z'_\gamma} I(z_d, z_m)dz_d dz_m \\ & + \int_{-\infty}^{z_\alpha}\int_{z'_\gamma}^{\infty} I(z_d, z_m)dz_d dz_m \\ = & \int_{-\infty}^{\infty}\int_{z'_\gamma}^{\infty} I(z_d, z_m)dz_d dz_m \end{aligned} \tag{B.34}$$

so it is sufficient to prove that the integral is positive when the range (z_α, ∞) is replaced with $(-\infty, \infty)$. Similar arguments can be used when $b_1 - b_0 > 0$, in which case it is sufficient to prove positivity when $z_\gamma = 0$.

This enables z_d (or z_m) to be integrated out, namely reducing to showing that

$$\begin{aligned} & \int_{z_\beta}^{\infty}\int_{z_\gamma}^{\infty} N_{\left(\begin{smallmatrix}1 & \rho_{sm}\\ \rho_{sm} & 1\end{smallmatrix}\right)}((z_s\, z_m)^t) - N_{\left(\begin{smallmatrix}1 & \rho_{rm}\\ \rho_{rm} & 1\end{smallmatrix}\right)}((z_s\, z_m)^t)dz_m dz_s > 0 \\ \Leftrightarrow & \int_{z_\beta}^{\infty} \frac{1}{2\pi}\exp\left(\frac{1}{2}z_s^2\right)\left(\Phi\left(\frac{\rho_{sm}z_s - z_\gamma}{1-\rho_{sm}^2}\right) - \Phi\left(\frac{\rho_{rm}z_s - z_\gamma}{1-\rho_{rm}^2}\right)\right) > 0 \end{aligned} \tag{B.35}$$

The second part of the integrand is monotonically increasing in z_s as $\rho_{sm} > \rho_{rm}$. Thus the integral is minimised as $z_\beta \to -\infty$, at which the value is $\Phi(z_\gamma)$, which is positive.

B.1.3 SNPs with aberrant allele frequency in one group

$R_B < R_A$ for SNPs with aberrance in C_1

If SNPs have aberrant MAF in C_1 only, we have $E(z_d) = \zeta_d \neq 0$, $E(z_m) = \zeta_m \neq 0$ and $E(z_s) = E(z_r) = 0$. As noted in the main text, as $\zeta_d \to 0$, $R_B, R_A \to P_0$ (equation 4.2 in the main paper) and

$$\begin{aligned}\lim_{\zeta_d\to\infty} R_B = & \lim_{\zeta_d\to\infty}\left(\int_{z_\alpha-\zeta_d}^{\infty}\int_{z_{\beta^*}}^{\infty}\int_{z_\gamma-\zeta_m}^{\infty} N_{\Sigma_B}\left((z_d\, z_s\, z_m)^t\right)dz_s dz_m dz_d \right. \\ & \left. + \int_{z_\alpha+\zeta_d}^{\infty}\int_{z_{\beta^*}}^{\infty}\int_{z_\gamma+\zeta_m}^{\infty} N_{\Sigma_B}\left((z_d\, z_s\, z_m)^t\right)dz_s dz_m dz_d\right) \\ = & \Phi(-z^*_\beta) = \frac{\beta^*}{2} \end{aligned} \tag{B.36}$$

and similarly, $R_A \to \frac{\beta}{2}$, $R_B \to \frac{\beta^*}{2}$ as $\zeta_d \to \pm\infty$, with $\beta^* < \beta$ as shown above. For β^* defined by 4.3 in the main paper, I show here that $R_A > R_B$ for all ζ_d. For the more general definition of β^* (equation 4.2 in the main paper), the inequality $R_B < R_A$ may not hold for all ζ_d. However, in practice, the inequality holds for almost all ζ_d and any deviation is small and near $\zeta_d = 0$.

Define the shorthand $N_\rho(x,y)$ as the value at (x,y) of the bivariate normal PDF with mean $\binom{0}{0}$ and variance $\left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)$, and $\operatorname{erfc}(x) = 2\left(1-\Phi(\sqrt{2}x)\right)$ as the complementary error function. In this section, ρ refers to ρ_{ds}.

Consider the value $R_A - R_B$ as a function of ζ_d. I will show that the partial derivative $\frac{\delta}{\delta\zeta_d}(R_A - R_B)$ is zero only when $\zeta_d = 0$. Because $R_A - R_B$ passes through the origin and is asymptotically positive in both directions, it is positive for all $\zeta_d \neq 0$. We have

$$\begin{aligned} R_A - R_B = &\left(\int_{z_\beta}^{\infty}\int_{z_\alpha-\zeta_d}^{\infty} N_0(x,y)dxdy - \int_{z_{\beta^*}}^{\infty}\int_{z_\alpha-\zeta_d}^{\infty} N_\rho(x,y)dxdy\right) \\ &+ \left(\int_{-\infty}^{-z_\beta}\int_{-\infty}^{-z_\alpha+\zeta_d} N_0(x,y)dxdy - \int_{-\infty}^{-z_{\beta^*}}\int_{-\infty}^{-z_\alpha+\zeta_d} N_\rho(x,y)dxdy\right) \end{aligned} \tag{B.37}$$

$$\begin{aligned} \frac{\delta}{\delta\zeta_d}(R_A - R_B) = &\left(\int_{z_\beta}^{\infty}\frac{\delta}{\delta\zeta_d}\int_{z_\alpha-\zeta_d}^{\infty} N_0(x,y)dxdy - \int_{z_{\beta^*}}^{\infty}\frac{\delta}{\delta\zeta_d}\int_{z_\alpha-\zeta_d}^{\infty} N_\rho(x,y)dxdy\right) \\ &+ \left(\int_{z_\beta}^{\infty}\frac{\delta}{\delta\zeta_d}\int_{z_\alpha+\zeta_d}^{\infty} N_0(x,y)dxdy - \int_{z_{\beta^*}}^{\infty}\frac{\delta}{\delta\zeta_d}\int_{z_\alpha+\zeta_d}^{\infty} N_\rho(x,y)dxdy\right) \\ = &\frac{1}{2\sqrt{2\pi}}\operatorname{erfc}\left(\frac{z_\beta}{\sqrt{2}}\right)\left(e^{-\frac{1}{2}(\zeta_d-z_\alpha)^2} - e^{-\frac{1}{2}(\zeta_d+z_\alpha)^2}\right) \\ -\frac{1}{2\sqrt{2\pi}}&\left(e^{-\frac{1}{2}(\zeta_d-z_\alpha)^2}\operatorname{erfc}\left(\frac{z_{\beta^*}+\rho(\zeta_d-z_\alpha)}{\sqrt{2(1-\rho^2)}}\right) - e^{-\frac{1}{2}(\zeta_d+z_\alpha)^2}\operatorname{erfc}\left(\frac{z_{\beta^*}-\rho(\zeta_d+z_\alpha)}{\sqrt{2(1-\rho^2)}}\right)\right) \end{aligned}$$

Showing that $\frac{\delta}{\delta\zeta_d}(R_A - R_B) > 0$ when $\zeta_d > 0$ is equivalent to showing that $(a-b) - (pa - qb) > 0$ where

$$\begin{aligned} a &= e^{-\frac{1}{2}(\zeta_d-z_\alpha)^2} & b &= e^{-\frac{1}{2}(\zeta_d+z_\alpha)^2} \\ p &= \frac{\operatorname{erfc}\left(\frac{z_{\beta^*}+\rho(\zeta_d-z_\alpha)}{\sqrt{2(1-\rho^2)}}\right)}{\operatorname{erfc}\left(\frac{z_\beta}{2}\right)} & q &= \frac{\operatorname{erfc}\left(\frac{z_{\beta^*}+\rho(\zeta_d-z_\alpha)}{\sqrt{2(1-\rho^2)}}\right)}{\operatorname{erfc}\left(\frac{z_\beta}{2}\right)} \end{aligned} \tag{B.38}$$

Since $(\zeta_d - z_a)^2 < (\zeta_d + z_a)^2$ for $\zeta_d > 0$, we have $a > b$. Because erfc is strictly decreasing, we have $p < q$. Because $\frac{\delta p}{\delta \zeta_d} < 0$, we have

$$p < \frac{\text{erfc}\left(\frac{z_{\beta^*} - z_\alpha)}{\sqrt{2(1-\rho^2)}}\right)}{\text{erfc}\left(\frac{z_\beta}{2}\right)} < 1 \tag{B.39}$$

where the second inequality arises because $z_{\beta^*} > \sqrt{1-\rho^2} z_\beta + \rho z_\alpha$. Thus $pa - qb < pa - pb = p(a-b) < a - b$, and $\frac{\delta}{\delta \zeta_d}(R_A - R_B) > 0$ as required.

B.1.4 Upper bound on $R_B - R_A$ with aberrance in C_1'

For SNPs with aberrance in C_1', we have $E(z_d) = 0$, $E(z_r) = \zeta_r \neq 0$, $E(z_s) = \zeta_s \neq 0$ and $E(z_m) = \zeta_m \neq 0$. As above $R_A, R_B \to P_0$ as $\zeta_r \to 0$, and by similar arguments to the section above, $R_A, R_B \to \frac{\alpha}{2}$ as $\zeta_r \to \pm\infty$.

For β^* defined as per equation 4.3 in the main paper, it is possible to derive an approximate (asymptotically accurate) upper bound on $R_B - R_A$, corresponding to the most serious increase in FPR. The approach is not readily applied to the general definition of β^*, but again the difference is typically small in practice.

To first order

$$\frac{\zeta_s}{\zeta_r} = \frac{\left(\frac{\mu_1' - \mu_0'}{SE(m_1' - m_0')}\right)}{\left(\frac{\mu_1' - \mu_0}{SE\left(m_1' - \frac{m_0 n_0 + m_0' n_0'}{n_0 + n_0'}\right)}\right)} = \sqrt{\frac{(n_0 + n_0')(n_0' + n_1'}{n_0'(n_0 + n_0' + n_1')}} \stackrel{\text{def}}{=} k \tag{B.40}$$

Now

$$\begin{aligned} R_B - R_A = &\left(\int_{z_{\beta^*} - \zeta_s}^{\infty} \int_{z_\alpha}^{\infty} N_\rho(x,y)dxdy - \int_{z_\beta - \zeta_r}^{\infty} \int_{z_\alpha}^{\infty} N_0(x,y)dxdy\right) \\ &+ \left(\int_{z_{\beta^*} + \zeta_s}^{\infty} \int_{z_\alpha}^{\infty} N_\rho(x,y)dxdy - \int_{z_\beta + \zeta_r}^{\infty} \int_{z_\alpha}^{\infty} N_0(x,y)dxdy\right) \end{aligned} \tag{B.41}$$

Define z_r^+, z_r^- such that

$$\begin{aligned}\int_{z_r^-}^{\infty}\int_{z_\alpha}^{\infty} N_0(x,y)dxdy &= \int_{z_{\beta^*}-\zeta_s}^{\infty}\int_{z_\alpha}^{\infty} N_\rho(x,y)dxdy \\ \int_{z_r^+}^{\infty}\int_{z_\alpha}^{\infty} N_0(x,y)dxdy &= \int_{z_{\beta^*}+\zeta_s}^{\infty}\int_{z_\alpha}^{\infty} N_\rho(x,y)dxdy\end{aligned} \tag{B.42}$$

From equation B.14 in Appendix B.1.2, we have $z_{\beta^*}-\zeta_s \approx \sqrt{1-\rho^2}z_r^- - \rho z_\alpha$ and $z_{\beta^*}+\zeta_s \approx \sqrt{1-\rho^2}z_r^+ - \rho z_\alpha$.

Noting that $\int_a^\infty\int_b^\infty N_0(x,y)dxdy = \Phi(-a)\Phi(-b)$ and $\Phi(x) = 1-\Phi(-x)$ we now have

$$R_B - R_A = \Phi(-z_\alpha)\left(\Phi(z_\beta-\zeta_r)-\Phi(z_r^-)+\Phi(z_\beta+\zeta_r)-\Phi(z_r^+)\right) \tag{B.43}$$

Applying the identity $\Phi(-z_\alpha) = \frac{\alpha}{2}$ and approximations $z_\beta^* \approx \sqrt{1-\rho^2}z_\beta + \rho z_\alpha$, $\zeta_s \approx k\zeta_r$, yields

$$\begin{aligned}z_r^- &\approx \frac{z_{\beta^*}-\zeta_s+\rho z_\alpha}{\sqrt{1-\rho^2}} \approx z_\beta - \frac{k}{\sqrt{1-\rho^2}}\zeta_r \\ z_r^+ &\approx z_\beta + \frac{k_1}{\sqrt{1-\rho^2}}z_0'\end{aligned} \tag{B.44}$$

$$R_B - R_A \approx \frac{\alpha}{2}\left(\Phi\left(z_\beta - \frac{k}{\sqrt{1-\rho^2}}\zeta_r\right) - \Phi(z_\beta-\zeta_r) + \Phi\left(z_\beta + \frac{k}{\sqrt{1-\rho^2}}\zeta_r\right) - \Phi(z_\beta+\zeta_r)\right) \tag{B.45}$$

Considered as a function of ζ_r, the value $R_B - R_A$ will be 0 at $\zeta_r = 0$ and tend to 0 as $\zeta_r \to \pm\infty$. It will be maximised approximately at the points where $\Phi(z_\beta-\zeta_r)$ or $\Phi(z_\beta+\zeta_r)$ are changing most rapidly; that is, $\zeta_r = \pm z_\beta$. At $\zeta_r = z_\beta$, the contribution to the value $R_B - R_A$ from the difference $\Phi\left(z_\beta + \frac{k}{\sqrt{1-\rho^2}}\zeta_r\right) - \Phi(z_\beta+\zeta_r)$ is negligible (and similarly for the other difference when $\zeta_r = -z_\beta$). Using the first-order approximation for $\Phi(z_\beta-\zeta_r)$ about $\zeta_r = z_\beta$ yields

$$\max\left(R_B - R_A\right) \approx \frac{\alpha}{2\sqrt{2\pi}}\left(\frac{k}{\sqrt{1-\rho^2}} - 1\right)z_\beta \tag{B.46}$$

In general, this value is substantially less than α.

B.1.5 Aberrance in C_0'

For SNPs aberrant in C_0', again $E(z_d)=0$, $E(z_r)=\zeta_r\neq 0$, $E(z_s)=\zeta_s\neq 0$ and $E(z_m)=\zeta_m\neq 0$. As above $R_A,R_B\to P_0$ as $\zeta_r\to 0$, and $R_A,R_B\to\frac{\alpha}{2}$ as $\zeta_r\to\pm\infty$. In this case, R_B is typically less than R_A.

B.1.6 General aberrance in replication cohorts

For β^* defined according to 4.3 in the main paper, the increase in FPR R_B-R_A for method B for a SNP with aberrance in C_1' is generally smaller than the decrease in FPR R_A-R_B for a SNP with a similarly-sized aberrance in C_0', in that the integral of the difference over the range of ζ_r is generally smaller in the former.

Define k as in the section above and $k_1=\frac{\zeta_s}{\zeta_r}\big|C_0'$ aberrant $=\sqrt{\frac{n_0'(n_0'+n_1')}{(n_0+n_0')(n_0+n_0'+n_1')}}$ similarly. Now for C_0' aberrant

$$R_A-R_B\approx\frac{\alpha}{2}\left(\Phi(z_\beta-\zeta_r)-\Phi\left(z_\beta-\frac{k_1}{\sqrt{1-\rho^2}}\zeta_r\right)+\Phi(z_\beta+\zeta_r)-\Phi\left(z_\beta+\frac{k_1}{\sqrt{1-\rho^2}}\zeta_r\right)\right) \tag{B.47}$$

Since $\int_0^x\Phi(z)dz=x\Phi(x)+\frac{1}{\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}-1\right)$, we have

$$\int_0^\infty(\Phi(h-z)-\Phi(h-kz))\,dz=\left(1-\frac{1}{k}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}h^2}+h\Phi(h)\right) \tag{B.48}$$

$$\int_0^\infty(\Phi(h+z)-\Phi(h+kz))\,dz=\left(1-\frac{1}{k}\right)\left(-\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}h^2}+h\Phi(-h)\right) \tag{B.49}$$

Thus with aberrant C_0'

$$\int_0^\infty(R_A-R_B)\,d\zeta_r=\frac{\alpha}{2}\left(1-\frac{\sqrt{1-\rho^2}}{k_1}\right)z_\beta \tag{B.50}$$

Comparing R_A and R_B under the two aberrance scenarios with the same ζ_d

$$\frac{\int_0^\infty(R_A-R_B)\,d\zeta_d\ [C_0'\text{ aberrant}]}{\int_0^\infty(R_B-R_A)\,d\zeta_d\ [C_1'\text{ aberrant}]}=\frac{1-\frac{\sqrt{1-\rho^2}}{k_1}}{\frac{\sqrt{1-\rho^2}}{k}-1} \tag{B.51}$$

For this to be > 1, a necessary condition is $\left(1 - \frac{\sqrt{1-\rho^2}}{k_2}\right) > \left(\frac{\sqrt{1-\rho^2}}{k_1} - 1\right)$ From the definitions of ρ_{ds} (Appendix B.1.1), k (equation B.40) and k_1, this is equivalent to

$$\sqrt{\frac{n_0+n_0'+n_1'}{n_0'+n_1'}}\sqrt{1-\frac{n_0 n_1 n_1'}{(n_0+n_0')(n_0+n_1)(n_0+n_0'+n_1')}}\left(\sqrt{\frac{n_0'}{n_0+n_0'}}+\sqrt{\frac{n_0+n_0'}{n_0}}\right) > 2$$

The final term in this product is of the form $x + \frac{1}{x}$ so is greater than 2. A sufficient condition is thus

$$\begin{aligned}\frac{n_0+n_0'+n_1'}{n_0'+n_1'}\left(1-\frac{n_0 n_1 n_1'}{(n_0+n_0')(n_0+n_1)(n_0+n_0'+n_1')}\right) &\geq 1 \\ \iff n_0^2+n_0(n_0'+n_1)+n_1(n_0'-n_1') &\geq 0 \end{aligned} \tag{B.52}$$

which holds in most study designs.

B.2 Supplementary figures

Please see following page

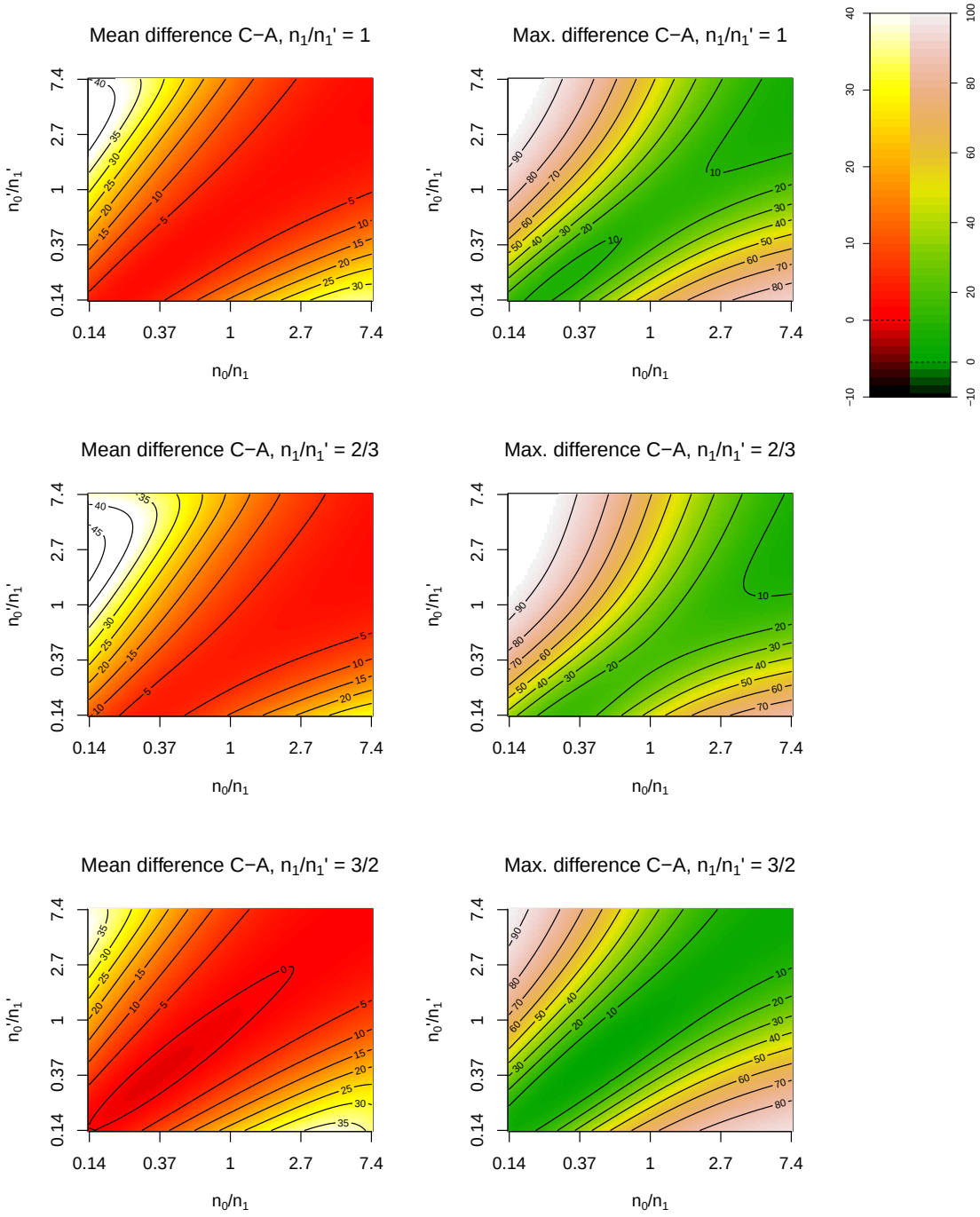


Fig. B.1 Power difference (%) between methods C and A. Mean power difference is taken as the integral of power difference between methods (see methods section) over $\mathbb{R}$ with respect to log-odds ratio. In all cases, 20 000 samples are used overall for a SNP with MAF 0.1, with cutoffs $\alpha = 5 \times 10^{-6}$, $\beta = 5 \times 10^{-4}$, $\gamma = 5 \times 10^{-8}$. Method C is almost universally more powerful.

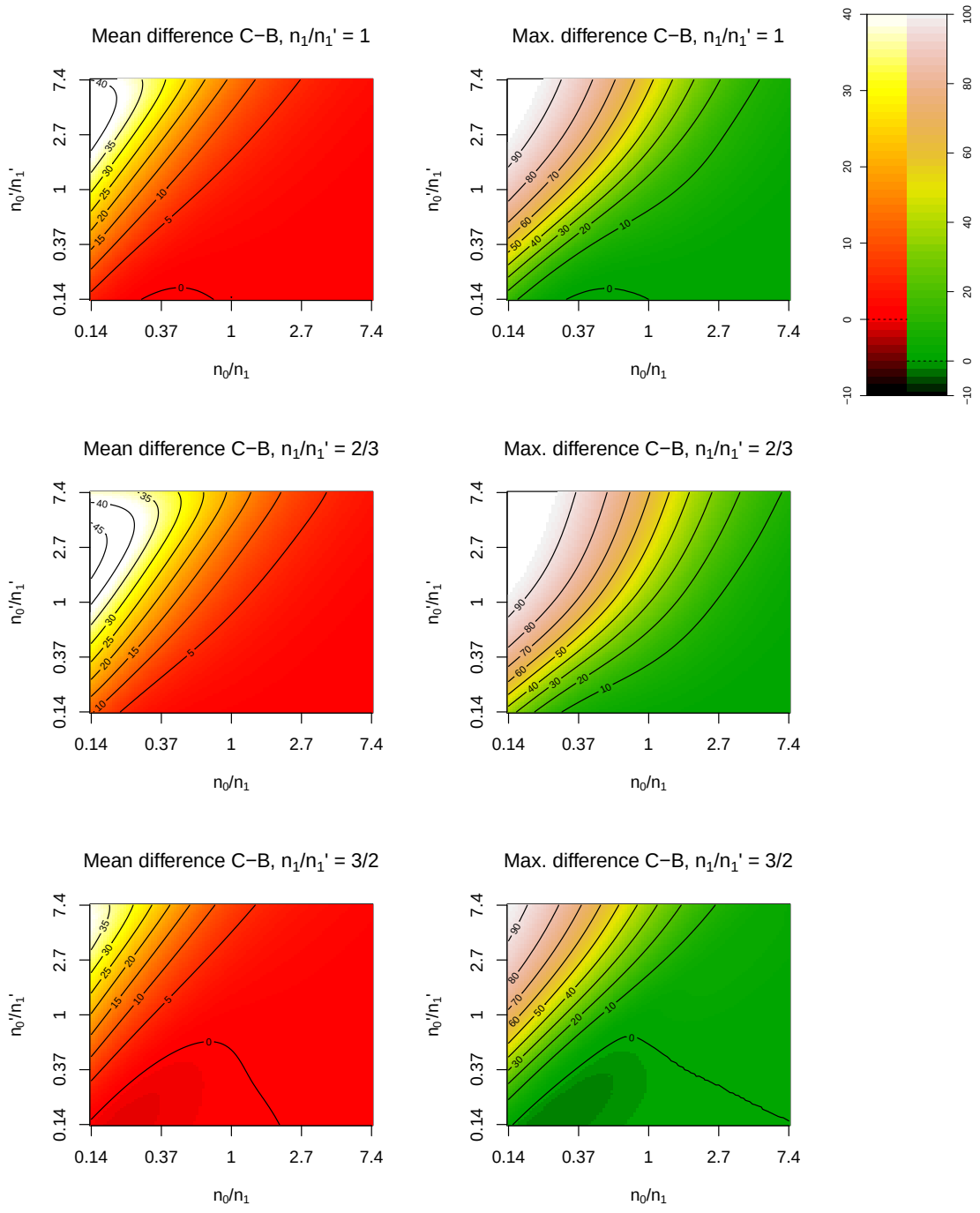


Fig. B.2 Power difference (%) between methods C and B. Mean power difference is taken as the integral of power difference between methods (see methods section) over $\mathbb{R}$ with respect to log-odds ratio. In all cases, 20 000 samples are used overall for a SNP with MAF 0.1, with cutoffs $\alpha = 5 \times 10^{-6}$, $\beta = 5 \times 10^{-4}$, $\gamma = 5 \times 10^{-8}$.

Appendix C

Supplementary note for chapter 5

C.1 Disease models in H_1 and H_0

I define 'differential causative pathology' (the alternative hypothesis, H_1) to mean that some subset of disease-associated variants have different population effect sizes in the case subgroups in question. My method tests against the null hypothesis H_0 that all disease associated variants have the same effect sizes in both subgroups. An equivalent formulation of H_0 is that the (possibly empty) sets of SNPs which have different minor allele frequencies in case and control groups and which have different minor allele frequencies in case subgroups are non-intersecting.

The multitude of potential causes for disease heterogeneity necessitate that both H_0 and H_1 encompass a range of such causes. I list several below, with illustration in appendix D.1, table D.1. I define the 'genetic architecture' of a trait as a set of variants and corresponding effect sizes (log-odds ratios or asymptotically similar statistics) between populations with and without the trait. In general, most effect sizes are zero or negligibly small.

C.1.1 Disease models in H_1

The simplest model of disease heterogeneity in H_1 is the scenario in which some variants are associated with one case subgroup, but not the other. For such a variant the effect size in one subgroup is zero, and in the other nonzero. This would be expected to arise if some of the pathological processes giving rise to the disease were specific to one case subgroup.

A second potential model in H_1 is when the same variants are associated with both subgroups, but the relative effect sizes differ. This may arise in a situation where pathological processes differ in relative impact between subgroups. For instance, if two pathological

processes may lead to a disease of interest, and one process is likely to occur during the neonatal period while the another is likely to occur during adolescence, a division of a case group into neonatal-onset and adolescent-onset would likely show variants associated with the first process as being more important in the first subgroup, and variants associated with the second process as being more important in the second, although the set of associated variants may be the same in both subgroups. The scenario may also arise if the cases can be split into subgroups like those described in the first paragraph, but the subgrouping criterion is only an approximation to this split.

A third model is when the same variants are associated with both subgroups with but where the effect sizes in one subgroup are a constant factor larger than in the other subgroup. This corresponds to differential heritability between subgroups, with the same pathological processes present. In a liability threshold model where some environmental variable has an additive effect with genetic risk, we would expect that defining subgroups based on the environmental variable would lead to this scenario (figure C.1). In this case, the environment modulates the effect of the genetic risk. As an example, under the assumption that a dietary risk factor has an additive effect with genetic risk factors in type 2 diabetes, a disease subgroup with the dietary risk factor would be expected to have lower disease heritability than a subgroup without it.

C.1.2 Disease models in H_0

Under H_0, all disease associated variants have the same effect size in both subgroups. This may take the form of an absence of any systematic genetic difference between case subgroups, in which case the population allelic frequencies of disease-associated SNPs, and hence the effect sizes of such SNPs between controls and each case subgroup, are equal.

Hypothesis H_0 also allows the presence of genetic differences between subgroups at different SNPs to those associated with the disease. This may be particularly prominent if variation in the disease depends on how the disease process acts on different individual physiologies, in which case genetic variation between subgroups is at different SNPs to those involved in disease causality.

C.1.3 Subgrouping by a risk factor

Partitioning a case group by a known disease risk factor may lead to subgroupings in either H_0 or H_1 dependent on the interaction between the genetic and environmental risk factors. If the risk factor on which the subgrouping is based has a multiplicative effect on disease risk with

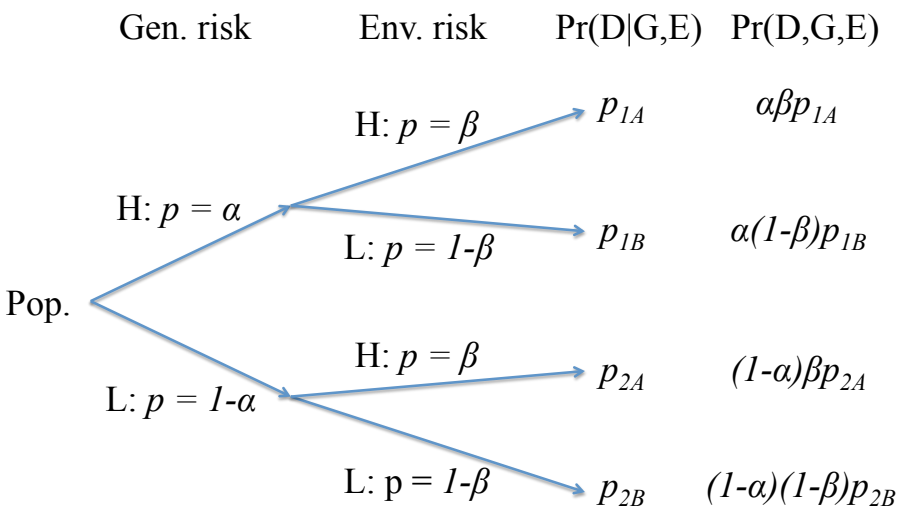


Fig. C.1 In a simplistic disease model, consider two levels of genetic risk G with frequencies α, $1-\alpha$ and an independent two-level environmental risk factor E with frequencies β, $1-\beta$, and a disease D. In cases with the environmental risk factor, we would expect the ratio of high-genetic risk to low-genetic risk cases to be $\frac{\alpha}{1-\alpha}\frac{p_{1A}}{p_{2A}}$, and in cases without, $\frac{\alpha}{1-\alpha}\frac{p_{1B}}{p_{2B}}$. Assume we define subgroups based on the environmental risk factor. If the risk factor has a multiplicative effect on $Pr(D|G,E)$, so $\frac{p_{1A}}{p_{2A}} = \frac{p_{1B}}{p_{2B}}$, the prevalences of genetic risk groups are identical in the groups, and the heritability of D is the same. If the effect of the environmental risk factor on $Pr(D|G,E)$ changes with G, so the environmental risk factor modulates the genetic risk, this will not hold.

genetic factors, then we expect the subgrouping to be in H_0 (figure C.1). This may take the form of a binary risk factor: if a disease is triggered by an environmental event (for example, a particular mutation driven by environmental mutagens), with susceptibility to that event determined genetically (for instance, impaired ability to repair the mutation), conditioning on environment will not affect the distribution of genetic risk, and the subgrouping will be in H_0. The genetic risk may also be binary; for example, the development of a disease may require the knock-out of a particular cellular process, with the genetic risk for the disease solely involved in risk of the knock-out.

However, deviation from a locally multiplicative model can also lead to a subgrouping in H_1. One instance this may occur is if disease risk approaches 1. A current model of T1D pathogenesis requires the presence of an environmental insult to trigger genetic susceptibility ([Rodriguez-Calvo et al., 2016]), which could be expected to lead to a locally multiplicative relationship between age-at-diagnosis and genetic risk (figure C.2). However, if genetic risk can be high enough that some individuals are almost sure to get the disease, this will lead to the subgrouping being in H_1 - a potential reason for the observation regarding age-at-diagnosis in T1D in the main text.

Finally, cases may be sub-grouped according to non-causative clinical disease associations. Assume some binary clinical marker M has non-zero frequency in healthy individuals and has some set of associated genetic variants G_0. Let D be a genetically homogeneous disease with a set of associated variants G_1 such that $G_0 \cap G_1 = \emptyset$ and D (or a necessary precursor of D) probabilistically causes M to occur more often than in the general population. Then when we condition on case status (and hence any necessary precursors of D) the only variants which are associated with M-status in cases will be in G_0, and a subgrouping based on M will be in H_0, despite M being associated with D. If, however, subtypes of D with differential genetic basis induce M to different degrees, and hence M serves as an index of such subtypes of D, then a subgrouping of M will fall in H_1.

C.2 Distribution of Z scores

In this section, I define the test statistics (Z scores) used to characterise allelic differences between groups and describe the rationale for my probabilistic model.

I partition SNPs into three theoretical categories:

1. SNPs which are not associated with case/control status or case subgroup status

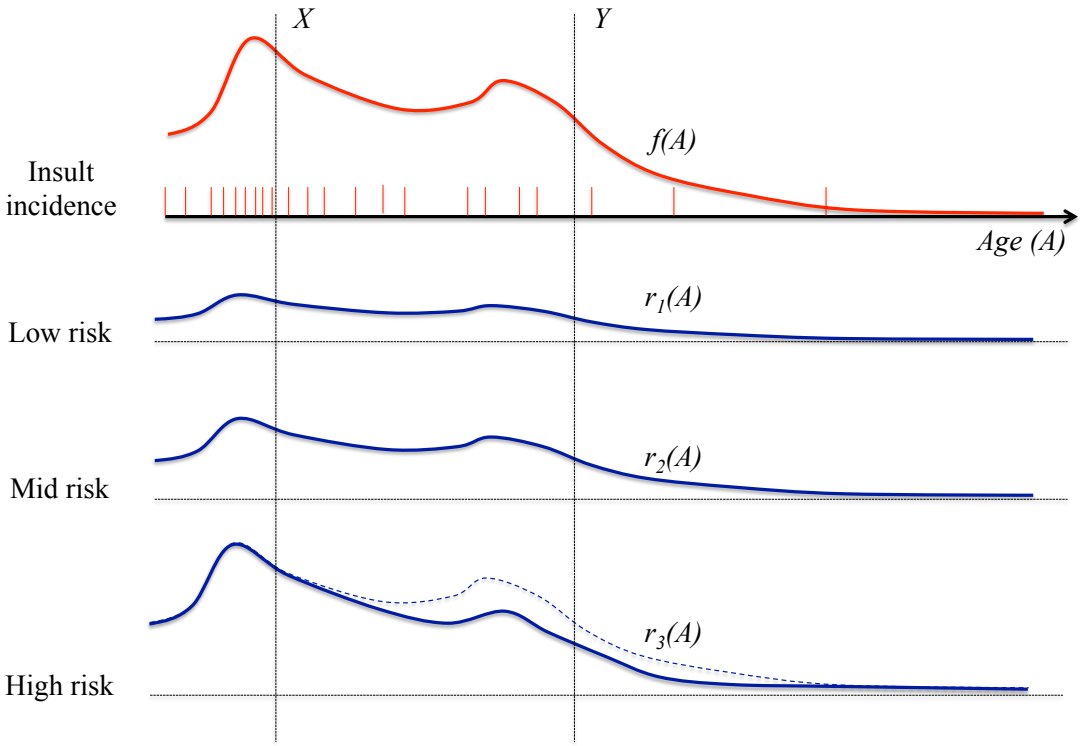


Fig. C.2 In a simplified model of incidence of type 1 diabetes or a similar autoimmune disease, I consider the disease to be triggered by an environmental 'insult'; for instance (eg, a viral illness) and three levels of genetic susceptibility to such insults. Denoting by $f(A)$ the density of such insults at age A (red vertical lines show a possible example for one individual), I expect that for individuals at low or moderate genetic risk the densities $r_1(A)$, $r_2(A)$ of disease incidence are proportional to $f(A)$, with lifetime risk $\int r_1(A)dA$, $\int r_2(A)dA$ respectively. The risk of disease at age A can be considered a product of $f(A)$ and a genetic risk score. In a high-risk group for a disease such as type 1 diabetes, it is possible that the lifetime risk $\int r_3(A)dA$ approaches 1, the high-risk group becomes 'saturated' with disease cases, and there are fewer non-affected individuals in the group at higher age groups, leading to a lower constant of proportionality with $f(A)$ at higher ages (dotted/solid lines). In the absence of the high-risk group, a subgrouping of patients into those with age-at-onset X and those with age-at-onset Y (vertical lines) would be expected to contain the same proportion of low- and mid- genetic risk samples in each subgroup, with correspondingly equal heritability of disease in each subgroup. With the high-risk group, the multiplicative effect of $f(A)$ on disease risk breaks down, inducing an environmental influence on the genetic risk, and changing the heritability between groups.

2. SNPs which are associated with the main phenotype but have the same effect size in both case subgroups

3. SNPs which are associated with the difference between case subgroups

I consider SNP effect sizes between subgroups and between cases and controls to be realisations of bivariate random variables, which have different distributions in each category.

C.2.1 Definitions

Unstratified groups

Let x be a random sample of size n_x from patient population X, and y a sample of size n_y from a population Y. Denote by m_x, m_y the allele frequencies of some SNP of interest in x and y, and by μ_x, μ_y the allele frequencies in X and Y. Assume for the moment that x and y are unbiased samples, so $\mu_x = E(m_x)$ and $\mu_y = E(m_y)$.

In general, I compute Z scores from GWAS -defined p-values P_{xy} using the formula

$$Z_{n_x,n_y}(m_x,m_y) = -\Phi^{-1}(P_{xy}/2)sign(m_x - m_y) \tag{C.1}$$

Although there are several ways in which a GWAS p-value may be computed, the resultant Z scores all have several common asymptotic properties. In general, I assume a Z score $Z_{n_x,n_y}(m_x,m_y)$ is a smooth function of allele frequencies m_x, m_y, n_x, n_y with the following properties

1. For fixed observed overall allele frequency $\frac{n_x m_x + n_y m_y}{n_x + n_y}$, $Z_{n_x,n_y}(m_x,m_y)$ is monotonic to the allelic difference $m_x - m_y$

2. Under the null hypothesis $\mu_x = \mu_y$,

   (a) $E(Z_{n_x,n_y}(m_x,m_y)) = 0$

   (b) $\text{var}(Z_{n_x,n_y}(m_x,m_y)) = 1$

   (c) $Z_{n_x,n_y}(m_x,m_y) \to_d N(0,1)$ as $n_x, n_y \to \infty$

These properties imply that the first-order expansion of Z about $(m_x,m_y) = (\mu,\mu)$ is:

$$Z_{n_x,n_y}(m_x,m_y) = \sqrt{\frac{2n_x n_y}{n_x + n_y}} \frac{m_x - m_y}{\sqrt{\mu(1-\mu)}} + O\left((m_x - \mu)(m_y - \mu)\right) \tag{C.2}$$

since

$$\begin{aligned}\sqrt{2n_x}(m_x-\mu_x) &\to_d N(0,\mu_x(1-\mu_x))\\ \sqrt{2n_y}(m_y-\mu_y) &\to_d N(0,\mu_y(1-\mu_y))\end{aligned} \tag{C.3}$$

and if $\mu_x=\mu_y=\mu$

$$\sqrt{\frac{2n_xn_y}{n_x+n_y}}\frac{m_x-m_y}{\sqrt{\mu(1-\mu)}} \to_d N(0,1) \tag{C.4}$$

and only one linear function of m_x, m_y can be asymptotically $N(0,1)$.

If $\mu_x \neq \mu_y$ and

$$\lambda = \frac{\mu_x-\mu_y}{\sqrt{\frac{\mu_x(1-\mu_x)}{2n_x}+\frac{\mu_y(1-\mu_y)}{2n_y}}} \tag{C.5}$$

remains finite as $n_x, n_y \to \infty$, we have

$$\begin{aligned}Z_{n_x,n_y}(m_x,m_y) &\approx \frac{m_x-m_y}{\sqrt{\frac{m_x(1-m_x)}{2n_x}+\frac{m_y(1-m_y)}{2n_y}}}\\ &= \frac{(m_x-m_y)-(\mu_x-\mu_y)}{\sqrt{\frac{m_x(1-m_x)}{2n_x}+\frac{m_y(1-m_y)}{2n_y}}}+\frac{\mu_x-\mu_y}{\sqrt{\frac{m_x(1-m_x)}{2n_x}+\frac{m_y(1-m_y)}{2n_y}}}\\ &\to_d N(0,1)+\lambda\\ &= N(\lambda,1)\end{aligned} \tag{C.6}$$

For a randomly chosen SNP, let μ_c be the population allele frequency (AF) in controls, and μ_1, μ_2 the population AFs in case subgroups 1 and 2 respectively, for the same allele. Define ν as the relative prevalence of subgroup 1 and $1-\nu$ as the relative prevalence of subgroup 2. The population AF across all cases is $\mu_{12}=\nu\mu_1+(1-\nu)\mu_2$.

Denote by m_c, m_1, m_2 the corresponding observed AFs in a study with n_c, n_1, n_2 controls and samples in subgroup 1 and subgroup 2 respectively. Define $m_{12}=\frac{n_1m_1+n_2m_2}{n_1+n_2}$ as the AF in the whole case group and $n_{12}=n_1+n_2$. Assume that $\frac{n_1}{n_1+n_2}\approx\nu$; that is, the case group is an unbiased sample of the case population. I later describe how this assumption can be relaxed.

The values Z_a and Z_d are defined as

$$Z_d = Z_{n_1,n_2}(m_1, m_2) \tag{C.7}$$

$$Z_a = Z_{n_1+n_2,n_c}(m_{12}, m_c) \tag{C.8}$$

where, as in the main paper, n_c, n_1, n_2 refer to cohort sizes and m_c, m_1, m_2 to observed allele frequencies in controls, subgroup 1 and subgroup 2 respectively, and m_{12} is the observed allele frequency across all cases (subgroups 1 and 2).

Adjustment for strata

If the distribution of some categorical variable (for example, country of origin) associated with allele frequency varies systematically between x and y, stratification may be needed when computing GWAS p-values. This may mean that $E(m_x) \neq E(m_y)$, even if the expected allele frequency is the same in x and y in each stratum. I proceed in a similar way to the approach in appendix B in determining correlation between Z scores with shared controls.

Assume x is divided into k strata $1..k$, and let n_x^1, n_x^2,..., n_x^k be the number of samples, m_x^1, m_x^2,..., m_x^k the observed allele frequencies and μ_x^1, μ_x^2,..., μ_x^k the expected allele frequencies for a SNP of interest in each stratum (and analagously for y).

Assume the Z score $Z_{\{n_x\},\{n_y\}}(\{m_x\},\{m_y\})$ in this case is a smooth function of $\{n_x^i\},\{n_y^i\},\{m_x^i\},\{m_y^i\}$, which has a first-order expansion about μ^1, μ^2, ... μ^k of the form

$$Z_{\{n_x\},\{n_y\}}(\{m_x\},\{m_y\}) = \frac{1}{\sqrt{\sum_{i\in 1..k} k_i^2 \frac{n_x^i+n_y^i}{2n_x^i n_y^i}\mu^i(1-\mu^i)}} \sum_{i\in 1..k} k_i(m_x^i - m_y^i) + O\left(\sum_{i\in 1..k} k_i(m_x^i - m_y^i)^2\right) \tag{C.9}$$

$$\approx \frac{\sum k_i(m_x^i - m_y^i)}{\sqrt{\mathrm{var}\left(\sum k_i(m_x^i - m_y^i) | \mu_x^i = \mu_y^i\right)}} + O\left(\sum k_i(m_x^i - m_y^i)^2\right) \tag{C.10}$$

where coefficients k_i depend only on the values $\{n_x\}$, $\{n_y\}$. For example, if the Cochran-Mantel-Haenszel test is used, $k_i = \frac{2n_x^i n_y^i}{n_x+n_y}$.

Using analogous definitions to those under subheading 'Unstratified groups', we now define

$$Z_d = Z_{\{n_1\},\{n_2\}}(\{m_1\},\{m_2\}) \tag{C.11}$$

$$Z_a = Z_{\{n_1+n_2\},\{n_c\}}(\{m_{12}\},\{m_c\}) \tag{C.12}$$

I term the coefficients of the allelic differences $m_1^i - m_2^i$, $m_{12}^i - m_c^i$ in the decomposition of Z_d and Z_a above as k_{di}, k_{ai} respectively.

Adjustment for covariates

If some continuous variable associated with the allele frequencies of some variant (for example, height) has a systematically different distribution in x and y, adjustment for covariates may be needed when computing GWAS p-values. In a similar approach to that in appendix B, section B.1.1 we set $G(i)$ as the numerical genotype of sample i (0,1,or 2) and w_i as the covariate value(s) for individual i. I consider w_i to be a sample from a random variable Z with PDF f_x in x and f_y in y.

Define the Z score $Z_{x,y}(\{G\},\{w\})$ in this case as a function of observed genotypes which permits a first-order expansion

$$Z_{x,y}(\{G\},\{w\}) = \frac{1}{\sqrt{\bar{m}(1-\bar{m})}}\left(\frac{1}{n_x}\sum_{i\in x} h_x(w_i)G(i) - \frac{1}{n_y}\sum_{j\in y} h_y(w_j)G(j)\right) \tag{C.13}$$

where h_x and h_y are functions of covariate scores, depending on the distribution of w in x and y, and parameter $\bar{m}$ is some measure of the overall allele frequency. The expected genotype of an individual may depend on their covariate value; for an individual i with covariate value(s) w_i in x set $g_x(w_i) = E(G(i))$, and set g_y similarly. Under the null hypothesis that the SNP of interest is not associated with x/y status, $g_x \equiv g_y$, and the expectation of Z must be 0. As in appendix B, section B.1.1, if the adjustment is to correct for every possible relation of the covariate to allelic frequencies (that is, have expectation 0 whenever $g_x = g_y$, whatever the form of g_x), we must have

$$h_x(w)f_x(w) = h_y(w)f_y(w) \tag{C.14}$$

for all covariate values w.

The sums of genotypes on the right of equation C.13 can be considered as 'effective' allele frequencies, and we define

$$\begin{aligned} m'_x &= \frac{1}{n_x}\sum_{i\in x} h_x(w_i)G(i) \\ m'_y &= \frac{1}{n_y}\sum_{j\in y} h_y(w_j)G(j) \end{aligned} \tag{C.15}$$

with expected values μ'_x, μ'_y respectively. Define 'effective' sample sizes $n'_x = \frac{\mu'_x(1-\mu'_x)}{\mathrm{var}(m'_x)}$, $n'_y = \frac{\mu'_y(1-\mu'_y)}{\mathrm{var}(m'_y)}$ so that, like allele frequencies (and under appropriate assumptions on the forms of f_x, f_y, g_x, g_y):

$$\frac{m'_x - \mu'_x}{\sqrt{\frac{\mu'_x(1-\mu'_x)}{n'_x}}} \rightarrow_d N(0,1) \tag{C.16}$$

and similarly for m'_y.

Now define

$$Z_d = Z_{\text{case 1,case 2}}(\{G\},\{w\}) \tag{C.17}$$

$$Z_a = Z_{\text{cases,controls}}(\{G\},\{w\}) \tag{C.18}$$

C.2.2 Z_d and Z_a are conditionally independent in categories 1 and 2

Unstratified or stratified groups

For SNPs in categories 1 and 2, $\mu_1 = \mu_2$. Hence

$$\begin{aligned}
\mathrm{cov}(Z_d, Z_a) &\propto \mathrm{cov}(m_{12} - m_c, m_1 - m_2) \\
&= \mathrm{cov}\left(\frac{n_1 m_1 + n_2 m_2}{n_1 + n_2} - m_c, m_1 - m_2\right) \\
&= \frac{1}{n_1 + n_2}\left(\mathrm{cov}(n_1 m_1, m_1) - \mathrm{cov}(n_2 m_2, m_2)\right) \\
&= \frac{1}{n_1 + n_2}\left(\mu_1(1-\mu_1) - \mu_2(1-\mu_2)\right)
\end{aligned} \tag{C.19}$$

which is 0 in categories 1 and 2.

For stratified groups, the same holds for each stratum; that is, $\mathrm{cov}(m^i_{12} - m^i_c, m^i_1 - m^i_2) = 0$. The independence of Z_d and Z_a follows from the expression of Z_d and Z_a as proportional to sums of allelic differences within strata and independence of the allelic differences in each stratum.

Adjustment for covariates

If we are adjusting for covariates, then using the definitions above, and noting that $g_1 = g_2$ for SNPs in categories 1 and 2, and $h_1 f_1 = h_2 f_2$ we have (defining $\mathbb{D}(w)$ as the domain of w)

$$\begin{aligned}
\operatorname{cov}(Z_d, Z_a) &\propto \operatorname{cov}\left(\frac{1}{n_{12}}\sum_{j\in cases} h_{12}(w_j)G(j) - \frac{1}{n_c}\sum_{i\in ctl} h_c(w_i)G(i),\right. \\
&\qquad \left.\frac{1}{n_1}\sum_{i\in c1} h_1(w_i)G(i) - \frac{1}{n_2}\sum_{j\in c2} h_2(w_j)G(j)\right) \\
&\propto \frac{1}{n_1}\sum_{i\in c1} h_{12}(w_i)h_1(w_i)\operatorname{var}(G(i)) - \frac{1}{n_2}\sum_{j\in c2} h_{12}(w_j)h_2(w_j)\operatorname{var}(G(j)) \\
&\to \int_{\mathbb{D}(w)} h_{12}(w)\left(h_1(w)f_1(w) - h_2(w)f_2(w)\right)g_1(w)(1-g_1(w))dw \\
&= 0 \qquad \text{(C.20)}
\end{aligned}$$

The simplifications are possible because genotypes vary independently in each cohort; that is, $\sum_{i\in ctl} h_c(w_i)G(i) \perp\!\!\!\perp \sum_{i\in c1} h_1(w_i)G(i)$, $\sum_{j\in c2} h_2(w_j)G(j)$, and $\sum_{i\in c1} h_{12}(w_i)G(i) \perp\!\!\!\perp \sum_{j\in c2} h_2(w_j)G(j)$, $\sum_{j\in c2} h_{12}(w_j)G(j) \perp\!\!\!\perp \sum_{i\in c1} h_1(w_i)G(i)$.

C.2.3 SNPs in category 3

Under H_0, SNPs in category 3 have the same allele frequency in cases and controls but different population allele frequencies between subgroups. Such a set may arise if subgrouping is based on some partially genetically-determined trait which is independent of the main phenotype has the same prevalence in case and control groups. An example may be subgroups defined by heterogeneity in treatment response arising only from individual pharmacokinetic variation. Under this assumption, the marginal variance of the joint distribution of Z_d, Z_a in the direction of Z_a is 1, and Z_d, Z_a are uncorrelated.

Under H_1 we expect SNPs in category 3 to be associated both with case/control status and with subgroup status. It is therefore expected that the marginal variances of the joint distribution to be greater than 1 in both the Z_a and Z_d directions, and possible correlation/anticorrelation between Z_a and Z_d.

Define $\zeta(\mu_x, \mu_y)$ as the population normalised log odds ratio between μ_x and μ_y:

$$\begin{aligned}\zeta(\mu_x, \mu_y) &= \sqrt{\bar{\mu}(1-\bar{\mu})}\log\left(\frac{\mu_x(1-\mu_y)}{\mu_x(1-\mu_y)}\right) \\ &= \frac{\mu_x - \mu_y}{\sqrt{\bar{\mu}(1-\bar{\mu})}} + O\left((\mu_x - \mu_y)^2\right)\end{aligned} \tag{C.21}$$

where $\bar{\mu} = \frac{1}{2}(\mu_x + \mu_y)$. For a set of SNPs of interest, we consider μ_1, μ_2, μ_c to be distributed such that $\zeta_d = \zeta(\mu_1, \mu_2)$ and $\zeta_a = \zeta(\mu_{12}, \mu_c)$ can be considered to be random variables with joint PDF:

$$F_{\sigma_a^2, \sigma_d^2, \rho_0} = \frac{1}{2}\left(N_{\left(\binom{0}{0}, \begin{pmatrix}\sigma_d^2 & \rho_0 \\ \rho_0 & \sigma_a^2\end{pmatrix}\right)} + N_{\left(\binom{0}{0}, \begin{pmatrix}\sigma_d^2 & -\rho_0 \\ -\rho_0 & \sigma_a^2\end{pmatrix}\right)}\right) \tag{C.22}$$

with σ_d, σ_a, and ρ_0 independent of n_1, n_2, n_c. Under H_0, $\sigma_a = 0$ (same MAFs in cases/controls) and $\rho_0 = 0$. I assume that ζ_d and ζ_a are conserved across strata and covariates.

Unstratified groups

Combining equation C.2 and expanding about $\bar{\mu}$:

$$\begin{aligned}\zeta(\mu_x, \mu_y) &= \frac{\mu_x - \mu_y}{\sqrt{\bar{\mu}(1-\bar{\mu})}} + O\left((\mu_x - \mu_y)^2\right) \\ &\approx \sqrt{\frac{n_x + n_y}{2n_x n_y}} Z_{n_x, n_y}(\mu_x, \mu_y)\end{aligned} \tag{C.23}$$

so defining $\bar{\mu}_d = \frac{1}{2}(\mu_1 + \mu_2)$ and $\bar{\mu}_a = \frac{1}{2}(\mu_{12} + \mu_c)$, we note (defining c_a and c_d):

$$\begin{aligned}E(Z_d|\bar{\mu}_d, \zeta_d) &= E(Z_d|\mu_1, \mu_2) \\ &= Z_{n_1, n_2}(\mu_1, \mu_2) \\ &= \sqrt{\frac{2n_1 n_2}{n_1 + n_2}}\zeta_d \\ &\stackrel{\text{def}}{=} c_d\,\zeta_d \\ E(Z_a|\bar{\mu}_a, \zeta_a) &= \sqrt{\frac{2n_{12} n_c}{n_{12} + n_c}}\zeta_a \\ &\stackrel{\text{def}}{=} c_a\,\zeta_a\end{aligned} \tag{C.24}$$

Set $\mu = (\mu_1, \mu_2, \mu_c)$. Since m_1, m_2 and m_c are conditionally independent given μ we have

$$\begin{aligned}
\operatorname{cor}(Z_a, Z_d|\mu) &= \operatorname{cor}(m_{12} - m_c, m_1 - m_2|\mu) \\
&= \frac{\operatorname{cov}(\frac{n_1 m_1 + n_2 m_2}{n_1+n_2} - m_c, m_1 - m_2|\mu)}{\sigma(m_s - m_c|\mu)\sigma(m_1 - m_2|\mu)} \\
&= \frac{\operatorname{cov}(n_1 m_1, m_1|\mu) - \operatorname{cov}(n_2 m_2, m_2|\mu)}{(n_1+n_2)\sigma(m_s - m_c|\mu)\sigma(m_1 - m_2|\mu)} \\
&= \frac{\mu_1(1-\mu_1) - \mu_2(1-\mu_2)}{(n_1+n_2)\sigma(m_s - m_c|\mu)\sigma(m_1 - m_2|\mu)} \\
&\approx 0
\end{aligned}$$

From equation C.6, $\operatorname{var}(Z_d|\mu_1, \mu_2) = \operatorname{var}(Z_a|\mu_1, \mu_2, \mu_c) = 1$. Thus approximately:

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} |\zeta_d, \zeta_a \sim N\left(\begin{pmatrix} c_d\,\zeta_d \\ c_a\,\zeta_a \end{pmatrix}, I_2 \right) \tag{C.25}$$

and the PDF of $(Z_a\ Z_d)^t$ at (x, y) has value

$$\begin{aligned}
&\iint_{\mathbb{R}^2} N_{(c_d\zeta_d\, c_a\zeta_a)^t, I_2}(x,y) F_{\sigma_a^2, \sigma_d^2, \rho_0}(\zeta_d, \zeta_a)\, d\zeta_d\, d\zeta_a \\
=&F_{1+c_a^2\sigma_a^2, 1+c_d^2\sigma_d^2, c_a c_d \rho_0}(x,y) \\
=&\frac{1}{2}\Big(N_{\left(\binom{0}{0}, \begin{pmatrix} 1+c_d^2\zeta_d^2 & c_a c_d\rho_0 \\ c_a c_d\rho_0 & 1+c_a^2\zeta_a^2 \end{pmatrix}\right)}(x,y) + N_{\left(\binom{0}{0}, \begin{pmatrix} 1+c_d^2\zeta_d^2 & -c_a c_d\rho_0 \\ -c_a c_d\rho_0 & 1+c_a^2\zeta_a^2 \end{pmatrix}\right)}(x,y)\Big)
\end{aligned} \tag{C.26}$$

which is a symmetric two-Gaussian distribution. Under H_0, the marginal variance in the direction of Z_a (fitted σ_3^2) is 1, and the covariance between Z_d and Z_a is zero.

Adjustment for strata

For stratified groups, we assume ζ_a and ζ_d are conserved across strata, and set $\bar{\mu}_d^i = \frac{1}{2}(\mu_1^i + \mu_2^i)$, $\bar{\mu}_a^i = \frac{1}{2}(\mu_{12}^i + \mu_c^i)$, k_{di} as the coefficient of $m_1^i - m_2^i$ in the first-order expansion of Z_d

(equation C.9), and k_{ai} as the coefficient of $m^i_{12} - m^i_c$ in the first-order expansion of Z_a to find

$$\begin{aligned}
E(Z_d|\{\bar{\mu}_d\},\zeta_d) &= E(Z_d|\{\bar{\mu}_1\},\{\bar{\mu}_2\}) \\
&\approx \frac{1}{\sqrt{\sum\limits_{i\in 1..k} k^2_{di}\frac{n^i_1+n^i_2}{2n^i_1n^i_2}\bar{\mu}^i_d(1-\bar{\mu}^i_d)}} \sum_{i\in 1..k} k_{di}(\mu^i_1-\mu^i_2) \\
&\approx \frac{\sum k_{di}\sqrt{\bar{\mu}^i_d(1-\bar{\mu}^i_d)}}{\sqrt{\sum k^2_{di}\frac{n^i_1+n^i_2}{2n^i_1n^i_2}\bar{\mu}^i_d(1-\bar{\mu}^i_d)}}\zeta_d \\
&\approx \frac{\sum k_{di}}{\sqrt{\sum k^2_{di}\frac{n^i_1+n^i_2}{2n^i_1n^i_2}}}\zeta_d \\
&\stackrel{\text{def}}{=} c'_d\zeta_d
\end{aligned} \tag{C.27}$$

and

$$\begin{aligned}
E(Z_a|\{\bar{\mu}^1_a,\bar{\mu}^2_a,...,\bar{\mu}^k_a\},\zeta_a) &\approx \frac{\sum k_{ai}}{\sqrt{\sum k^2_{ai}\frac{n^i_{12}+n^i_c}{2n^i_{12}n^i_c}}}\zeta_a \\
&\stackrel{\text{def}}{=} c'_a\zeta_a
\end{aligned} \tag{C.28}$$

assuming that for most SNPs the values $\bar{\mu}^i_d$, $\bar{\mu}^i_a$ do not differ markedly across strata. If the Cochran-Mantel-Haenszel test is used,

$$\begin{aligned}
c'_d &= \sqrt{\sum k_{di}} \\
&= \sqrt{\sum \frac{2n^i_1n^i_2}{n^i_1+n^i_2}} \\
c'_a &= \sqrt{\sum \frac{2n^i_{12}n^i_c}{n^i_{12}+n^i_c}}
\end{aligned} \tag{C.29}$$

and the PDF of Z_d, Z_a is then as for equation C.26 with c'_d, c'_a in place of c_d, c_a.

Adjustment for covariates

The expression for $Z_{x,y}(\{G\},\{w\})$ can be rewritten as:

$$Z_{x,y}(\{G\},\{w\}) = \frac{1}{\sqrt{\bar{m}(1-\bar{m})}}(m'_x - m'_y) \tag{C.30}$$

As above, I assume that whatever the dependence of $E(G(i))$ on covariates, the allelic difference between case subgroups is constant, so $g_1(w) - g_2(w)$ is constant for all w. Define the analogue of $\zeta(\mu_x, \mu_y)$ given covariate(s) w

$$\zeta(\mu_x, \mu_y)|w = \sqrt{\bar{\mu}(w)(1-\bar{\mu}(w))} \log\left(\frac{\mu_x'(w)(1-\mu_y(w))}{\mu_x(w)(1-\mu_y(w))}\right) \tag{C.31}$$

and define c_d'', c_a'' analogously to the previous section:

$$c_d'' = \frac{E\{Z_d\}}{\zeta_d}$$
$$c_a'' = \frac{E\{Z_a\}}{\zeta_a}$$

The joint distribution of Z_d and Z_a is then given by the analogue of equation C.26 with appropriate analogues of c_d, c_a.

C.2.4 Unequal subgroup prevalences

Motivation

The criteria by which subgroups are defined may have a different distribution in the population than in the case group, with the consequence that the disease subtype corresponding to one of the subgroups may be oversampled relative to its true prevalence in the population.

This leads to inaccuracies in the inferred genetic architecture recovered from a case-control study (ie, a typical GWAS), which may take the form of false-positive associations. If there exist variants which differentiate subgroups, oversampling of one subgroup will bias the the observed overall variant effect sizes toward the effect size in the oversampled subgroup, even if the variants are un-associated with the phenotype overall.

In serious cases, this could lead to false identification of variants associated only with subgroup status as associated with the disease as a whole. For example, a GWAS on rheumatoid arthritis (RA) in which the case group had a high prevalence of obesity may identify purely obesity-associated variants as RA-associated.

For stratified and covariate-adjusted analyses, the equivalent problem is failure of population subgroup prevalences to match study subgroup prevalences within each strata or across covariates. This could be a result of ascertainment bias; different geographic locations could report different frequencies of disease subtypes due to differences in clinic specialities.

As well as affecting conventional GWAS analyses, I show below that subgroup oversampling can cause false-positives in my test. I show a modification to the method to account for this.

Behaviour of standard approach

I demonstrate the effect of mismatched sample and population subgroup frequencies in the scenario where no strata or covariates are used. The extension to the generalised cases is similar.

Assume that in the disease population, the 'true' prevalences of subgroups 1 and 2 are ν, $1-\nu$, and define $\mu_{12} = \nu\mu_1 + (1-\nu)\mu_2$ as the underlying MAF across all cases in the population. In the hypothesis test to compute P_a, the hypothesis $H_a : \mu_c = \frac{n_1 m_1 + n_2 m_2}{n_1+n_2}$ is not equivalent to $H : \mu_c = \mu_{12}$.

Since $E(m_{12}) = E\left(\frac{n_1 m_1 + n_2 m_2}{n_1+n_2}\right) = \frac{n_1\mu_1 + n_2\mu_2}{n_1+n_2} \neq \mu_{12}$, equation C.25 becomes

$$\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} | \zeta_d, \zeta_a \sim N\left(\begin{pmatrix} Z_{n_1,n_2}(\mu_1,\mu_2) \\ Z_{n_{12},n_c}(\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2}, \mu_c) \end{pmatrix}, I_2\right) \tag{C.32}$$

Now

$$\begin{aligned} Z_{n_{12},n_c}\left(\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2}, \mu_c\right) &\approx \frac{c_a}{\sqrt{\bar{\mu}(1-\bar{\mu})}}\left(\frac{n_1\mu_1+n_2\mu_2}{n_1+n_2} - \mu_c\right) \\ &= \frac{c_a}{\sqrt{\bar{\mu}(1-\bar{\mu})}}\left((\mu_{12}-\mu_c) + \left(\frac{n_1}{n_1+n_2} - \nu\right)(\mu_1-\mu_2)\right) \\ &\approx c_a(\zeta_a + k\zeta_d) \end{aligned} \tag{C.33}$$

where $k = (\frac{n_1}{n_1+n_2} - \nu)$, so the unconditional distribution of $(Z_a\ Z_d)^t$ in this case is given by

$$\begin{aligned} &\iint_{\mathbb{R}^2} N_{(c_d\zeta_d\ c_a(\zeta_a+c\zeta_d))^t, I_2}(x,y) F_{\sigma_a^2,\sigma_d^2,\rho_0}(\zeta_d,\zeta_a)\, d\zeta_d\, d\zeta_a \\ &= \frac{1}{2}\left(N_{\left(\binom{0}{0},\Sigma_2\right)}(x,y) + N_{\left(\binom{0}{0},\Sigma_3\right)}(x,y)\right) \end{aligned} \tag{C.34}$$

where

$$\begin{aligned} \Sigma_2 &= \begin{pmatrix} 1+c_d^2\zeta_d^2 & c_a c_d(\rho_0 + k\zeta_d^2) \\ c_a c_d(\rho_0 + k\zeta_d^2) & 1+c_a^2(\zeta_a^2 + k^2\zeta_d^2 + 2k\rho_0) \end{pmatrix} \\ \Sigma_3 &= \begin{pmatrix} 1+c_d^2\zeta_d^2 & c_a c_d(-\rho_0 + k\zeta_d^2) \\ c_a c_d(-\rho_0 + k\zeta_d^2) & 1+c_a^2(\zeta_a^2 + k^2\zeta_d^2 - 2k\rho_0) \end{pmatrix} \end{aligned} \tag{C.35}$$

Distribution C.34 consists of the sum of two Gaussians which are not mirror images in the x and y axes. Conceptually, the aberrance between prevalences of subgroups in the population and in the study induces a bias in Z_a toward either Z_1 or Z_2, whichever is comparatively over-represented in the study compared to the population.

This effect is demonstrated in figure C.3, with simulated data and approximate distribution as per C.34. As the discrepancy between the relative proportions grows, the distributions precess around the origin. Importantly, under H_0 ($\sigma_a = 0$, $\rho_0 = 0$) the distribution of Z_d, Z_a will not satisfy $\sigma_3 = 1$, $\rho = 0$, and my standard approach is inappropriate.

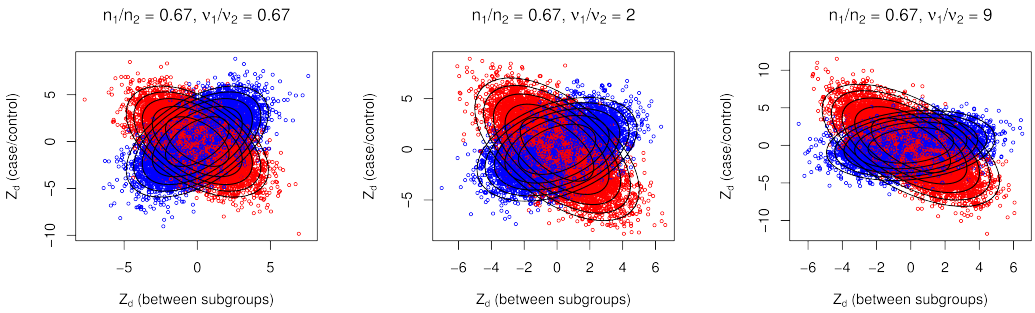


Fig. C.3 Distribution of (Z_a, Z_d) for SNPs in category 3 when observed subgroup frequency (n_1/n_2) does not match underlying subgroup frequency in the population $(\nu_1/\nu_2 = \nu/(1-\nu))$. Red and blue points correspond to the two Gaussian distributions comprising the underlying distribution of effect sizes. Contour lines of distributions are shown. Note the precession in the axes of the distributions as the difference between ν_1/ν_2 and n_1/n_2 increases, and loss of symmetry when $\nu_1/\nu_2 \neq n_1/n_2$

Adaptation

If the true proportion of case subgroups in the population are known, the problem of over-sampled subgroups can be overcome by a recalculation of Z_a. The problem broadly arises because the expected value of the observed allele frequency in cases, $E(m_{12})$, is different from the true allele frequency μ_{12} in cases in the population, for SNPs in category 3.

This can be addressed by using an unbiased estimate of the true population allele frequency $m'_{12} = \nu m_1 + (1-\nu)m_2$ in place of m_{12}. The resultant Z score, Z'_a, is obtained by adjusting Z_a by subtracting a multiple of Z_d:

$$Z'_a = \frac{1}{\sqrt{1+\beta^2}}(Z_a - \beta Z_d) \tag{C.36}$$

so, given a between-subgroup effect size ζ_d, $\text{var}(Z'_a|\zeta_d) = 1$. I choose β so that $E(Z'_a) = 0$ for SNPs in category 3 (see below). The adjustment leads to systematic nonzero covariance between Z_d and Z'_a.

Z_a and Z_d are independent conditioned on ζ_a, ζ_d and $\bar{\mu}$. Thus under H_0 and conditioning on $\bar{\mu}$

$$\begin{aligned}
\text{cov}(Z'_a, Z_d|\zeta_a, \zeta_d) &= \frac{1}{\sqrt{1+\beta^2}} E(Z_d(Z_a - \beta Z_d)|\zeta_a, \zeta_d) \\
&= \frac{1}{\sqrt{1+\beta^2}} \left(E(Z_d Z_a|\zeta_a, \zeta_d) - \beta E(Z_d^2|\zeta_a, \zeta_d)\right) \\
&= \frac{-\beta}{\sqrt{1+\beta^2}} \text{var}(Z_d^2|\zeta_d) \\
&= \frac{-\beta}{\sqrt{1+\beta^2}}
\end{aligned} \tag{C.37}$$

and because ζ_d and ζ_a are independent under H_0, $\text{cov}(Z'_a, Z_d) = \frac{-\beta}{\sqrt{1+\beta^2}}$ in every category. I denote this consistent covariance by ρ_c

Hence the overall model for Z_d, Z_a changes to

$$\begin{aligned}
PDF_{Z_d,Z_a|\Theta}(d,a) =& \pi_1 N_{\left(\begin{smallmatrix} 1 & \rho_c \\ \rho_c & 1 \end{smallmatrix}\right)}(d,a) && \text{(category 1)} \\
&+ \pi_2 N_{\left(\begin{smallmatrix} 1 & \rho_c \\ \rho_c & \sigma_2^2 \end{smallmatrix}\right)}(d,a) && \text{(category 2)} \\
&+ \pi_3 \left(\frac{1}{2} N_{\left(\begin{smallmatrix} \tau^2 & \rho+\rho_c \\ \rho+\rho_c & \sigma_3^2 \end{smallmatrix}\right)}(d,a) + \frac{1}{2} N_{\left(\begin{smallmatrix} \tau^2 & -\rho+\rho_c \\ -\rho+\rho_c & \sigma_3^2 \end{smallmatrix}\right)}(d,a) \right) && \text{(category 3)}
\end{aligned} \tag{C.38}$$

where, under H_0, $\rho = 0$ and $\sigma_3 = 1$. This requires a slight modification of the fitting algorithm. My R package at https://github.com/jamesliley/subtest contains an implementation.

No stratification of covariates

If no strata nor covariates are used, set

$$\begin{aligned}
\beta &= \left(\frac{n_1}{n_1+n_2} - \nu \right) \frac{c_a}{c_d} \\
&\stackrel{\text{def}}{=} k \frac{c_a}{c_d}
\end{aligned} \tag{C.39}$$

recalling the definitions of c_a and c_d from equation C.24, and that ν is the proportion of cases of subgroup 1 in the population while $\frac{n_1}{n_1+n_2}$ is the proportion in the study. The value $k = \left(\frac{n_1}{n_1+n_2} - \nu\right)$ thus corresponds to the dissimilarity between subgroup prevalences in the case group and in the population.

Under H_0, for SNPs in category 3 we have

$$\begin{aligned} E(Z'_a|\zeta_d) &\propto E\left(Z_a - k\frac{c_a}{c_d}Z_d\right) \\ &= \frac{c_a}{\sqrt{\bar{m}(1-\bar{m})}}E\left(\left(\frac{n_1m_1+n_2m_2}{n_1+n_2} - m_c\right) - \left(\frac{n_1}{n_1+n_2} - \nu\right)(m_1-m_2)\right) \\ &= \frac{c_a}{\sqrt{\bar{m}(1-\bar{m})}}E\left(\nu m_1 + (1-\nu)m_2 - m_c\right) \\ &= 0 \end{aligned} \tag{C.40}$$

since $E(\nu m_1 + (1-\nu)m_2) = \nu\mu_1 + (1-\nu)\mu_2 = \mu_c = E(m_c)$ for all SNPs under H_0.

Adjustment for strata

In the equivalent adjustment for stratified groups, define

$$\beta = \sqrt{\frac{\sum k_{di}^2 \frac{n_1^i+n_2^i}{2n_1^in_2^i}}{\sum k_{ai}^2 \frac{n_{12}^i+n_c^i}{2n_{12}^in_c^i}}} \frac{\sum k_{ai}\left(\frac{n_1^i}{n_1^i-n_2^i} - \nu\right)}{\sum k_{di}} \tag{C.41}$$

so, assuming $\mu_1^i - \mu_2^i$ are conserved and $\bar{\mu}_a^i$, $\bar{\mu}_d^i$ are close to conserved across strata, and given $\bar{\mu}_a^i \approx \bar{\mu}_d^i | H_0$:

$$\begin{aligned}
E(Z_a' | H_0) &= \frac{\sum k_{ai}(\mu_{12}^i - \mu_c^i)}{\sqrt{\sum k_{ai}^2 \frac{n_{12}^i + n_c^i}{2n_{12}^i n_c^i} \bar{\mu}_a^i (1 - \bar{\mu}_a^i)}} + \beta \frac{\sum k_{di}(\mu_1^i - \mu_2^i)}{\sqrt{\sum k_{di}^2 \frac{n_1^i + n_2^i}{2n_1^i n_2^i} \bar{\mu}_d^i (1 - \bar{\mu}_d^i)}} \\
&\approx \frac{\sum k_{ai}(\mu_{12}^i - \mu_c^i)}{\sqrt{\bar{\mu}_a(1 - \bar{\mu}_a)} \sqrt{\sum k_{ai}^2 \frac{n_{12}^i + n_c^i}{2n_{12}^i n_c^i}}} + \beta \frac{\sum k_{di}(\mu_1^i - \mu_2^i)}{\sqrt{\bar{\mu}_d(1 - \bar{\mu}_d)} \sqrt{\sum k_{di}^2 \frac{n_1^i + n_2^i}{2n_1^i n_2^i}}} \\
&= \frac{1}{\sqrt{\bar{\mu}_a(1 - \bar{\mu}_a)} \sqrt{\sum k_{ai}^2 \frac{n_{12}^i + n_c^i}{2n_{12}^i n_c^i}}} \left(\sum k_{ai} \left(\frac{n_1^i \mu_1^i + n_2^i \mu_2^i}{n_1^i + n_2^i} - \mu_c^i \right) - \left(\frac{n_1^i}{n_1^i + n_2^i} - \nu \right) (\mu_1^i - \mu_2^i) \right) \\
&= \frac{1}{\sqrt{\bar{\mu}_a(1 - \bar{\mu}_a)} \sqrt{\sum k_{ai}^2 \frac{n_{12}^i + n_c^i}{2n_{12}^i n_c^i}}} \left(\sum k_{ai} ((\nu \mu_1^i + (1 - \nu)\mu_2^i) - \mu_c^i) \right) \\
&= 0
\end{aligned} \tag{C.42}$$

Adjustment for covariates

If covariates are used, define the functions h_{12}, h_1, f_1, f_2 in the same way as in subheading 'Adjustment for covariates' and set

$$\beta = \frac{\int_{\mathbb{D}(w)} h_{12}(w) \left((1 - \nu) f_1(w) - \nu f_2(w) \right) dw}{\int_{\mathbb{D}(w)} h_1(w) f_1(w) dw} \tag{C.43}$$

so

$$
\begin{aligned}
E(Z'_a|H_0) &\propto \sqrt{\bar{\mu}(1-\bar{\mu})}E(Z_a - \beta Z_d) \\
&= \left(\frac{1}{n_{12}} \sum_{i \in c1} h_{12}(w_i)G(i) + \frac{1}{n_{12}} \sum_{i \in c2} h_{12}(w_i)G(i) - \frac{1}{n_c} \sum_{i \in controls} h_c(w_i)G(i) \right) \\
&\quad - \beta \left(\frac{1}{n_1} \sum_{i \in c1} h_1(w_i)G(i) - \frac{1}{n_2} \sum_{i \in c2} h_2(w_i)G(i) \right) \\
&\to \int_{\mathbb{D}(w)} h_{12}(w) \left(f_1(w)g_1(w) + f_2(w)g_2(w) \right) - h_c(w)f_c(w)g_c(w)dw \\
&\quad - \beta \left(g_1(w) - g_2(w) \right) \int_{\mathbb{D}(w)} h_1(w)f_1(w)dw \\
&= \int_{\mathbb{D}(w)} h_{12}(w) \left(f_1(w) + f_2(w) \right) \left(\nu g_1(w) + (1-\nu)g_2(w) \right) - h_c(w)f_c(w)g_c(w)dw \\
&= \int_{\mathbb{D}(w)} h_{12}(w)f_{12}(w) \left(\nu g_1(w) + (1-\nu)g_2(w) - g_c(w) \right) dw \\
&= 0
\end{aligned} \tag{C.44}
$$

since $h_{12}f_{12} = h_c f_c$ and $h_1 f_1 h_2 f_2$ in the same way as under subheading 'Adjustment for covariates' in section C.2.1, $g_1 - g_2$ is constant by assumption, and the expected population genotypes at covariate value w are the same in cases $(\nu g_1(w) + (1-\nu)g_2(w))$ and controls $(g_c(w))$ under H_0.

C.3 Testing procedure

C.3.1 Algorithm

For testing a subgrouping S of interest, I use the following protocol (recalling definitions from chapter 5, section 5.4.2):

1. Compute Z_a scores between cases and controls
2. For the proposed subgrouping S
   (a) Compute scores Z_d^S corresponding to S,

(b) Fit parameters of full and null models

$$\theta_1^S = \arg\max_{\theta \in H_1} PL_{da}(Z_d^S, Z_a|\theta)$$
$$\theta_0^S = \arg\max_{\theta \in H_0} PL_{da}(Z_d^S, Z_a|\theta)$$

(c) Compute

$$uPLR = \log\{PL_{da}(Z_d^S, Z_a|\theta_1^S)\} - \log\{PL_{da}(Z_d^S, Z_a|\theta_0^S)\}$$

and adjusting factor

$$f(Z_a|\theta_1^S, \theta_0^S) = \log\{PL_a(Z_a|\theta_1^S)\} - \log\{PL_a(Z_a|\theta_0^S)\}$$

(d) Compute $PLR_S = uPLR - f(Z_a|\theta_1^S, \theta_0^S)$

3. Fix $\hat{\pi}_2$ and $\hat{\sigma}_2$ as per equation 5.16

4. For > 1000 random subgroups R of the case group

(a) Compute scores Z_d^R corresponding to R

(b) Fit parameters

$$\theta_1^R = \arg \max_{\theta \in H_1 | \pi_2 = \hat{\pi}_2, \sigma_2 = \hat{\sigma}_2} cPL(Z_d^R|Z_a, \theta)$$
$$\theta_0^R = \arg \max_{\theta \in H_0 | \pi_2 = \hat{\pi}_2, \sigma_2 = \hat{\sigma}_2} cPL(Z_d^R|Z_a, \theta)$$

noting the difference between maximising PL_{da} over all parameters in step 2b and maximising cPL over all parameters except π_2 and σ_2 here.

(c) Compute

$$cPLR_R = \log\{cPL(Z_d^R|Z_a, \theta_1^R)\} - \log\{cPL(Z_d^R|Z_a, \theta_0^R)\} \tag{C.45}$$

5. Estimate parameters γ, κ of the null distribution of $cPLR_R$ values (of the form $\gamma\left(\kappa\chi_1^2 + (1-\kappa)\chi_2^2\right)$)

6. Compute p-value for PLR_S by comparison to the distribution in step 5.

In summary, I compare an adjusted pseudo-log likelihood ratio for a subgrouping of interest to conditional pseudo-log likelihood ratios for randomly-chosen subgroupings. As discussed in section 5.4.2, an easier potential approach would be to compare PLR_S to PLR values PLR_R from random subgroups, but this is problematic.

Define one distribution F as 'majorising' another distribution G if, for $X \sim F$, $Y \sim G$, and all $x > 0$, we have $Pr(X > x) > Pr(Y > x)$ (with 'minorising' defined in the obvious way). If F majorises G and we wish to estimate the value $Pr(Y > x|X \sim G)$, then the value $Pr(X > x|X \sim F)$ is an overestimate. In the sense that $Pr(Y > x|X \sim G)$ is a p-value, the value $Pr(X > x|X \sim F)$ provides a conservative estimate of this p-value (that is, retaining control of the type-1 error rate).

For fixed parameters $\theta \in H_0$, the *uPLR* and *cPLR* both asymptotically have the same mixture-χ^2 distribution, which I will denote M. Because $PLR < uPLR$ (as f is positive), the distribution (asymptotic and empirical) of *uPLR* majorises the distribution of *PLR*, although the majorisation is slight when $\tau \gg 0$ (section C.4).

Unfortunately, parameter sets with $\tau = 1$ may not be in H_0 (see figure C.4) although such parameters correspond to no genetic differences between subgroups and are not of interest. Because of this, the distribution of $uPLR|\tau \approx 1$ may depart substantially from M and in general will majorise it (see figure C.5 in chapter 5). Departures from H_0 with $\tau = 1$ do not, however, tend to affect the cPLR, since it effectively 'removes' the contribution of Z_a alone to the PLR (section C.3.3, final two paragraphs). Indeed the cPLR tends to have a similar distribution both when $\tau = 1$ and when $\tau \gg 1$ (section C.4). The distribution of *PLR* with $\tau \approx 1$, by contrast, will minorise the distribution of *uPLR* and will generally also minorise the distribution of *cPLR* (see section C.3.3).

Since by resampling random subgroups, we can only sample the subspace of the parameter space with $\tau = 1$ and Z_a fixed, we are restricted to testing a *PLR* value from a subtyping of interest against the empirical distributions of one of $PLR|\tau = 1$, $uPLR|\tau = 1$, or $cPLR|\tau = 1$, bearing in mind that the subtyping of interest may have $\tau = 1$ or $\tau \gg 1$. Testing against the empirical distribution of $uPLR|\tau = 1$ will generally have little power when the value of τ in the subtyping of interest is $\gg 1$ since the distribution of $uPLR|\tau = 1$ substantially majorises the distribution of $uPLR|\tau \gg 1$. Testing against the empirical distribution of $PLR|\tau = 1$ may lead to loss of control of the type-1 error rate when the value of τ in the subtyping of interest is $\gg 1$, since the distribution of $PLR|\tau = 1$ minorises the distribution of $cPLR|\tau = 1$ and generally the distribution of $PLR|\tau > 1$. This leaves testing the *PLR* against the distribution of *cPLR*, which is slightly conservative when the value of τ in the subgrouping of interest is ≈ 1, and appropriate type-1 error control otherwise. This conservatism is in fact somewhat

desirable (see section 5.3, chapter 5), and hence this which was the approach I eventually took.

C.3.2 Use of *uPLR* for testing

A problem arises with the behaviour of the unadjusted pseudo-log likelihood ratio statistic $uPLR = \log\{PL_{da}(Z_d^S, Z_a|\Theta_1^S)\} - \log\{PL_{da}(Z_d^S, Z_a|\Theta_0^S)$ when the true value of τ (the marginal variance of Z_d in group 3) is near 1, corresponding to an absence of SNPs which differentiate subgroups.

If $\tau = 1$, there can be no differential genetic architecture between the subgroups, as there are no systematic genetic differences between them at all. However, the joint distribution of Z_d, Z_a may still be in H_1; if Z_a has an equally weighted three-Gaussian mixture distribution with variances $1, a^2, b^2$, and $Z_d \sim N(0,1)$, the true parameter values are $(\pi_2, \pi_3, \tau, \sigma_2, \sigma_3, \rho) = (\frac{1}{3}, \frac{1}{3}, 1, a, b, 0) \in H_1 \setminus H_0$ (figure C.4).

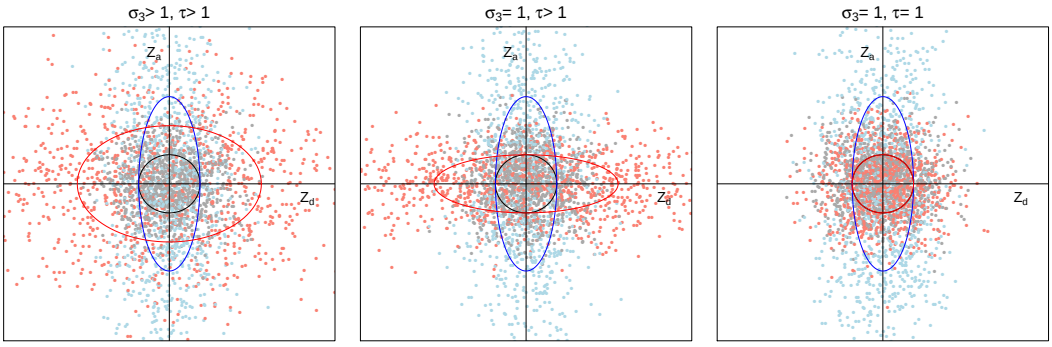


Fig. C.4 Potential for false positives when $\tau = 1$. Black/grey points and contours correspond to category 1, blue to category 2, and red/pink to category 3. Top two figures show potential distributions of Z_d, Z_a with $\sigma_3 > 1$; bottom two figures distributions with $\sigma_3 = 1$. A test based on the unadjusted pseudo-log likelihood ratio $uPLR = \log\{L(Z_d^S, Z_a|\Theta_1^S)\} - \log\{L(Z_d^S, Z_a|\Theta_0^S)\}$ will reject H_0 for both of the top two scenarios. However, we do not want to reject H_0 for the top right figure, in which $\tau = 1$ (no genetic difference between subgroups). This scenario is possible in real data, as the distribution of Z_a is only approximately normal and may more closely resemble a three-Gaussian mixture distribution (where components have variances σ_2^2, σ_3^2 and 1) than a two-Gaussian mixture distribution (where components have variances σ_1^2 and 1).

This problem is particularly prevalent in randomly-chosen subgroups, since $\tau = 1$ by assumption in this case. If the distribution of Z_d, Z_a from a test subgrouping is to be compared against corresponding distributions from random subgroupings, this problem must be addressed.

C.3.3 Rationale for approach

Consider the function

$$cPL(Z_d|Z_a,\theta) = \log\{PL_{da}(Z_d,Z_a|\theta)\} - \log\{PL(Z_a|\theta)\} \tag{C.46}$$

As per chapter 5, define

$$\widehat{\theta_1} = \arg\max_{\theta\in H_1} PL_{da}(Z_d,Z_a|\theta)$$
$$\widehat{\theta_0} = \arg\max_{\theta\in H_0} PL_{da}(Z_d,Z_a|\theta)$$

and given fixed values $\hat{\pi}_2$, $\hat{\sigma}_2$, define

$$\widehat{\theta_1^c} = \arg\max_{\theta\in H_1|\pi_2=\hat{\pi}_2,\sigma_2=\hat{\sigma}_2} cPL(Z_d|Z_a,\theta)$$
$$\widehat{\theta_0^c} = \arg\max_{\theta\in H_0|\pi_2=\hat{\pi}_2,\sigma_2=\hat{\sigma}_2} cPL(Z_d|Z_a,\theta)$$

Now

$$cPLR = \log\{cPL(Z_d|Z_a,\hat{\theta}_1^c)\} - \log\{cPL(Z_d|Z_a,\hat{\theta}_0^c)\}$$
$$PLR = \log\{cPL(Z_d|Z_a,\hat{\theta}_1)\} - \log\{cPL(Z_d|Z_a,\hat{\theta}_0)\}$$

Under H_0, with $\tau \approx 1$, I argue below (heuristically) that $\log\{cPL(Z_d|Z_a,\hat{\theta}_1^c) \geq \log\{cPL(Z_d|Z_a,\hat{\theta}_1)\}$ and $\log\{cPL(Z_d|Z_a,\hat{\theta}_0^c) \approx \log\{cPL(Z_d|Z_a,\hat{\theta}_0)\}$, so $cPLR \geq PLR$. This implies that for $\tau \approx 1$, we have $Pr(PLR < x|H_0) < Pr(cPLR < x|H_0)$. If $\tau \gg 1$ then both PLR and $cPLR$ have a mixture chi2 distribution under H_0, of the form in equation 5.2 in chapter 5. The scaling factor γ arises from LDAK weights, common to both PLR and $cPLR$, and the mixing parameter κ tends to be approximately $1/2$, so if $\tau \gg 1$, $Pr(PLR < x|H_0) \approx Pr(cPLR < x|H_0)$.

The value $PL_a(Z_a|\theta)$ can be considered an expected value with fixed values Z_a of $PL_{da}(Z_d,Z_a|\theta)$ over observations of Z_d. Since the parameters π_2, σ_2 characterise only the Z_a distribution, which is common to both $PL_a(Z_a|\theta)$ and $PL_{da}(Z_d,Z_a|\theta)$, we have

$$\frac{\partial}{\partial\pi_2}\log\{PL_{da}(Z_d,Z_a|\theta)\} \approx \frac{\partial}{\partial\pi_2}\{PL_a(Z_a|\theta)\}$$
$$\frac{\partial}{\partial\sigma_2}\log\{PL_{da}(Z_d,Z_a|\theta)\} \approx \frac{\partial}{\partial\sigma_2}\{PL_a(Z_a|\theta)\} \tag{C.47}$$

so $\frac{\partial cPL}{\partial \pi_2} \approx 0$ and $\frac{\partial cPL}{\partial \sigma_2} \approx 0$, and the value of cPL changes only slightly with changes in π_2, σ_2. Denote $\hat{\theta}_1' = \arg\max_{\theta \in H_1 | \pi_2 = \hat{\pi}_2, \sigma_2 = \hat{\sigma}_2} PL_{da}(Z_d, Z_a|\theta)$ (defined similarly to $\hat{\theta}_1$ but with π_2 and σ_2 fixed). Now [1]

$$cPL(Z_d|Z_a, \hat{\theta}_1) \approx cPL(Z_d|Z_a, \hat{\theta}_1') \tag{C.48}$$

$$\leq \max_{\theta \in H_1 | \pi_2 = \hat{\pi}_2, \sigma_2 = \hat{\sigma}_2} cPL((Z_d|Z_a, \theta)$$

$$= cPL((Z_d|Z_a, \hat{\theta}_1^c) \tag{C.49}$$

Now recalling the definition of $PL_a(Z_a|\theta)$:

$$PL_a(Z_a|\Theta) = \prod_{Z_a^{(i)} \in Z_a} \left(\pi_1 N_{0,1}(Z_a^{(i)}) + \pi_2 N_{0,\sigma_2^2}(Z_a^{(i)}) + \pi_3 N_{0,\sigma_3^2}(Z_a^{(i)}) \right) \tag{C.50}$$

we see that under H_0, the value of $PL_a(Z_a|\theta)$ is fixed if π_2 and σ_2 are fixed, since $\sigma_3 = 1$ and $\rho = 0$. Defining $\hat{\theta}_0'$ analogously to above, the constancy of PL_a implies that maximising PL_{da} is equivalent to maximising cPL under H_0, so $\hat{\theta}_0' = \hat{\theta}_0^c$ and hence

$$cPL(Z_d|Z_a, \hat{\theta}_0) \approx cPL(Z_d|Z_a, \hat{\theta}_0') = cPL((Z_d|Z_a, \hat{\theta}_0^c) \tag{C.51}$$

as required.

A rough explanation of this argument is that contributions to the unadjusted PLR can come from either the distribution of Z_a or the interaction between Z_a and Z_d, and inflation in the unadjusted PLR when $\tau = 1$ arise only from the former. If the former effect is large, the parameters in θ_1 will tend to be values which maximise the former effect, somewhat compromising the latter. By eliminating the former effect, using the adjustment, only this compromised contribution of the latter is allowed to contribute to the adjusted PLR. The compromise is why the distribution of $PLR > x|\tau \approx 1)$ approximately minorises the distribution of $cPLR > x|\tau \approx 1)$ (see section C.3.1). When τ is larger, the SNPs in categories 2 and 3 can be more easily distinguished, and the two effects above are more separate, so the compromise is lower. By contrast, the values which maximise the $cPLR$ effectively take into account the adjustment for Z_a (by maximising cPL rather than PL_{da}), and the compromise of the latter effect does not occur.

[1] Note that $\log\{PL_{da}(Z_d, Z_a|\hat{\theta}_1)\} - \log\{PL_{da}(Z_d, Z_a|\hat{\theta}_1')$ has a χ^2 distribution with 2 degrees of freedom under H_0 if $\hat{\pi}_2$ and $\hat{\sigma}_2$ are the true values of π_2 and σ_2, and hence it is not generally near 0. The argument applied in equation C.48 derived from equation C.47 is that $\log\{PL_{da}(Z_d, Z_a|\hat{\theta}_1)\} - \log\{PL_{da}(Z_d, Z_a|\hat{\theta}_1') \approx \log\{PL_a(Z_a|\hat{\theta}_1)\} - \log\{PL_a(Z_a|\hat{\theta}_1')$, from which approximation C.48 is reasonable.

As discussed in section **??**, if we were to use the adjusted *uPLR* to generate the null distribution using random subgroups, the majorisation of the observed distribution by the mixture-χ^2 may lead to loss of FDR control in test subgroups with $\tau > 1$. However, using the slightly anti-conservative distribution of *cPLR* to fit the null distribution overcomes this problem. Indeed, some conservatism is desirable when $\tau = 1$ as a double guard against rejecting H_0. The power of *cPLR* to reject H_0 is, however, somewhat lower than the power of the *PLR*, so I test using adjusted *uPLR* and fit the null distribution with *cPLR*.

C.4 Details of simulations

C.4.1 Simulations of random genotypes

Firstly, I simulated genotypes at independent SNPs to establish the distributions of *PLR* and *cPLR* under H_0 with $\tau = 1$ and $\tau > 1$.

I simulated the following scenarios:

1. (a) (Z_d, Z_a) under H_0 with $\tau = 1$
   (b) (Z_d, Z_a) under H_0 with τ allowed to vary
2. (Z_d, Z_a) under H_1

In each case, Z_a and Z_d were calculated from simulated genotypes at 5×10^4 independent autosomal SNPs in Hardy-Weinberg equilibrium. Because the sample size only affects PLR through the size of the fitted parameters (section C.4.3) I fixed the sample size at 2000 controls and 1000 cases of each subgroup and varied the underlying effect size distribution. Larger sample sizes correspond to larger deviations of underlying values of σ_2, σ_3, τ from 1 (table C.1).

For all simulations, I computed the uPLR and PLR (with adjustment $f(Z_a)$). For scenario 1a ($\tau = 1$, corresponding to random subgroups) I additionally computed the cPLR. Simulations 2 functioned as power calculations; the results from these are shown in the main text.

I tested over values of π_3 from $\{10^{-3}, 10^{-2}, 0.1, 0.2\}$. Values of σ_2, σ_3, τ were chosen corresponding to 97.5% quantiles of odds ratios in $\{1.5, 2, 2.5\}$ for case/control comparison (Z_a) or $\{1, 1.2, 1.5, 2\}$ for between-subgroups comparison (Z_d), table C.1.Values of ρ were chosen corresponding to correlations in $\{0, 0.1, 0.5\}$.

I compared the observed distributions of PLR from simulations 1a,1b with the observed distribution of cPLR from simulation 1a. Q-Q plots are shown in figure C.5. The distribution

	97.5% quantile of odds ratios			
n_1, n_2	1.2	1.5	2	2.5
500, 500	1.20	1.75	2.66	3.41
1000, 500	1.25	1.94	3.02	3.89
1000, 1000	1.36	2.27	3.62	4.71

Table C.1 Approximate expected standard deviations of observed Z scores for given odds-ratio distributions at various study sizes. For instance, if a study had 500 cases of each subgroup, and 95% of 'true' odds ratios (corresponding to population MAFs) for SNPs in category 3 were less than 1.5, the expected value of τ (the standard deviation of Z scores for SNPs in category 3) would be 2.66.

of cPLR agrees well with a mixture-χ^2 distribution, as does the distribution of PLR for simulation 1b. The distributions of PLR for simulations 1a,1b are minorised by the distribution of cPLR, more so for simulations 1a ($\tau = 1$), leading to a conservative test overall. Using *cPLR* to fit a null distribution, and using a significance cutoff $p < 0.05$, leads to a false-discovery rate of 0.048 (95% CI 0.039-0.059) in subgroups with $\tau > 1$ and 0.033 (95% CI 0.022-0.045) in subgroups with $\tau = 1$.

I also show the distribution of unadjusted PLR (uPLR) for simulations 1a and 1b. The distribution for 1a markedly majorises the mixture-χ^2 distribution, and has a very different distribution to that for 1b. Thus, if a test subgroup with $\tau >> 1$ was compared to random subgroups using unadjusted PLR, the test would have very low power to reject H_0. Finally, I plotted the estimated null distribution for all tests of real disease datasets, and found that the empirical distributions of cPLR from random subgroups agreed well with the proposed mixture χ^2 distribution (appendix D.2, figure D.5a, D.5b, D.5c).

C.4.2 Simulation on GWAS case group subgroups

To check the extensibility of these results to real data, I performed a similar set of simulations on data generated from subgroups of an ATD case group. In order to simulate scenarios in which $\tau > 1$, I selected subgroups for which groups of ≈ 50 SNPs differentiated subgroups without being associated with the disease in general.

Specifically, I repeatedly polled the overall dataset for sets of 2000 SNPs in linkage equilibrium, then clustered them hierarchically using a Euclidean distance metric. I then chose the first-appearing cluster of 50 SNPs, and hierarchically clustered the individuals in the case group according to a metric based on similarity across the 50 SNPs. When there were two clusters of individuals left, I denoted the two clusters as subgroup 1 and subgroup

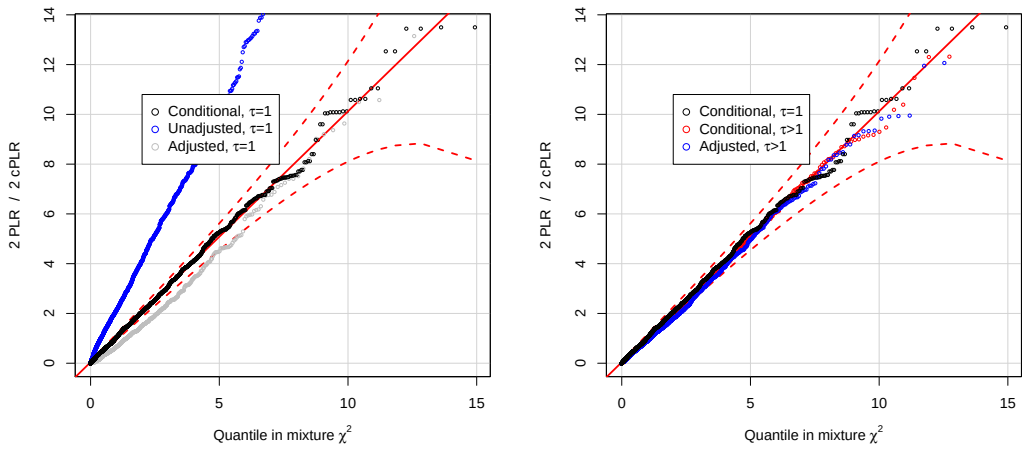


Fig. C.5 Q-Q plots comparing distributions of *PLR* and *cPLR* for subgroups based on simulated genotypes with a random variable distributed as $\frac{1}{2}(\chi_1^2 + \chi_2^2)$ (that is, $\gamma = \kappa = \frac{1}{2}$). In both plots, the black points correspond to conditional PLR (cPLR) values for 'random' subgroups ($\tau = 1$). The observed distribution is well-approximated by the asymptotic mixture-χ^2. The left-hand plot shows the distributions of unadjusted and adjusted PLR for subgroups with $\tau = 1$. The distribution of unadjusted PLR markedly majorises the mixture-χ^2, but the adjustment largely fixes this. The right-hand plot compares the distribution of cPLR for random subgroups with PLR for subgroups with $\tau > 1$. The distribution of cPLR is well-approximated by the mixture-χ^2 whether $\tau = 1$ (black) or $\tau > 1$ (red). In both plots, the distribution of cPLR and the mixture-χ^2 distribution slightly majorise the distribution of PLR, leading to a conservative test.

2. The mean resultant fitted value of τ was ≈ 5 and standard deviation of fitted values was ≈ 1.5.

For simulated subgroups with $\tau = 1$ (randomly chosen) and with $\tau > 1$ I computed *PLR* and *cPLR*. As for simulated genotypes, the resultant distributions showed good agreement with the proposed mixture-χ^2 distributions (figure C.6), with the approximation of the null distribution of *PLR* with the distribution of *cPLR* again leading to a conservative test, as expected. The type 1 error rate corresponding to $\alpha = 0.05$ was 0.52 (95% CI 0.043-0.061) in subgroups with $\tau > 1$ and 0.012 (95% CI 0.007-0.016) in subgroups with $\tau = 1$.

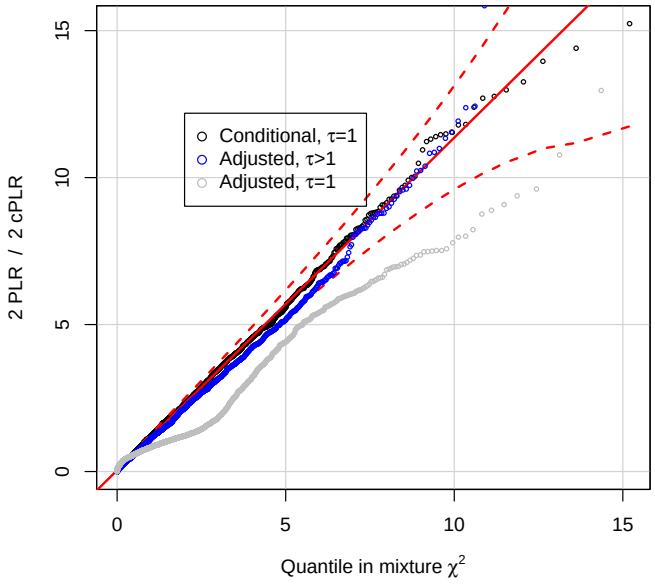


Fig. C.6 Comparison of distributions of *PLR* and *cPLR* for subgroups of an ATD case group, chosen so $\tau = 1$ or $\tau > 1$. The distribution of cPLR for random subgroups ($\tau = 1$) and the distribution of *PLR* for subgroups with $\tau >> 1$ are both well-approximated by a random variable distributed as $\frac{1}{2}(\chi^2_1 + \chi^2_2)$; red dashed lines show 99% pointwise confidence intervals. The distribution of *PLR* when $\tau = 1$ is minorised by the mixture-χ^2 leading to a conservative test if a subgroup with $\tau = 1$ is tested using PLR against the observed distribution of cPLR for random subgroups. Because $\tau = 1$ implies no genetic difference between subgroups, this is reasonable behaviour for the test.

C.4.3 Distributions of parameter values for simulation and power calculations

I assume a distribution of summary statistics parametrised by six variables: π_1, π_2, σ_2, σ_3, τ, and ρ (the value of π_3 is determined by π_1 and π_2). The space of all parameter values is too large to meaningfully assess performance of my test across it, so for each simulation, I draw the value of underlying parameters from sets of potential values chosen to reflect values which may arise in real data.

For a SNP S in two groups of size n_1, n_2, denote the population allele frequencies as μ_1, μ_2 and the corresponding observed allele frequencies as m_1, m_2. Set $\mu = \frac{\mu_1 n_1 + \mu_2 n_2}{n_1 + n_2}$ as the overall observed MAF, $r = \log\left(\frac{\mu_1(1-\mu_2)}{\mu_2(1-\mu_1)}\right)$ and $R = \log\left(\frac{m_1(1-m_2)}{m_2(1-m_1)}\right)$ as the 'underlying' and

observed log-odds ratios respectively. To first order

$$\begin{aligned} SE\{R\} &= \sqrt{\frac{1}{2m_1n_1} + \frac{1}{2(1-m_1)n_1} + \frac{1}{2m_2n_2} + \frac{1}{2(1-m_2)n_2}} \\ &\approx \sqrt{\frac{1}{2m(1-m)}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned} \tag{C.52}$$

The observed Z score is, to first order, $Z = \frac{R}{SE(R)}$. Now

$$\begin{aligned} E(Z|\mu, r) &\approx r\sqrt{\frac{2\mu(1-\mu)n_1n_2}{n_1+n_2}} \\ SD(Z|\mu, r) &\approx 1 \end{aligned} \tag{C.53}$$

Consider r as a $N(0, \sigma^2)$ random variable, and fix μ. Now, to first order

$$Z|\mu \sim N\left(0, 1 + \frac{2\mu(1-\mu)\sigma^2 n_1 n_2}{n_1+n_2}\right) \tag{C.54}$$

Assuming μ to have an approximately uniform distribution on $(0, 0.5]$, this gives

$$Z \sim N\left(0, 1 + \frac{\sigma^2 n_1 n_2}{3(n_1+n_2)}\right) \tag{C.55}$$

An interpretable description of the underlying odds-ratio distribution is the 0.975 quantile of 'true' odds ratios (approximately 2 standard deviations). If 97.5% of 'true' odds ratios r fall in $[1/\alpha, \alpha]$, then $\sigma \approx \frac{\log(\alpha)}{2}$ and the expected value of the corresponding observed standard deviation of Z (that is, σ_2, σ_3, or τ) is

$$\sqrt{1 + \frac{\log(\alpha)^2 n_1 n_2}{12(n_1+n_2)}} \tag{C.56}$$

Some examples are shown in table C.2:

		Study size (n_1/ n_2)					
α	*SD*	100/100	100/500	500/500	500/1000	1000/1000	2000/2000
1.1	0.05	1.02	1.03	1.09	1.12	1.17	1.32
1.2	0.09	1.07	1.11	1.30	1.39	1.54	1.94
1.3	0.13	1.13	1.22	1.56	1.71	1.97	2.60
1.5	0.20	1.30	1.46	2.10	2.36	2.80	3.83
2	0.35	1.73	2.08	3.31	3.79	4.58	6.41

Table C.2 Correspondence between odds-ratio distribution and standard deviation of observed Z score for various study sizes. Column α is the 97.5 % quantile of population odds-ratios for SNPs with non-zero effect sizes (approximately two standard deviations). Column *SD* is the corresponding standard deviation of the underlying log-odds ratio distribution (assumed to be normal). Entries in the table correspond to expected standard deviations of observed Z scores; that is, σ_2, σ_3 or τ. I allow different odds-ratio distributions between cases and controls for SNPs in categories 2 and 3 (corresponding to σ_2 and σ_3 respectively). For σ_2 or σ_3, n_1 is the number of cases and n_2 the number of controls; for τ; n_1 and n_2 are the number of cases in each disease subgroup.

C.5 Genetic correlation as an alternative to PLR test

C.5.1 Overview

The presence of genetic heterogeneity between disease subgroups could be tested for by adapting several known methods, although to my knowledge no specific method has yet been developed. One potential approach is to estimate the narrow-sense genetic correlation (r_g) across a set of SNPs between case/control traits of interest, either between Z scores derived from comparing the control group to each case subgroup, testing under the null hypothesis $r_g = 1$ (method 1); or between the familiar Z_a and Z_d, under the null hypothesis $r_g = 0$ (method 2).

This approach should have the advantage of characterising heterogeneity using a single widely-interpretable metric. However, both methods have, in this naive application, have multiple shortcomings which preclude their general use to subgroup testing. The most important of these are systematic false-positives arising in method 1, and false-negatives arising in method 2. I demonstrate this theoretically and in simulations. In addition, genetic correlation is a signed test statistic; genetic effects in the same direction contribute positively, and opposite directions contribute negatively, causing a loss of power in situations where pleiotropy between the phenotypes involves shared effects of both types. Finally, I found that

tests involving r_g were less powerful than the PLR in rejecting the null hypothesis in real genetic data (ATD; GD vs HT).

Genetic correlation is an estimate of the similarity in genetic basis of two traits. A useful formal definition is given by Bulik-Sullivan et al [Bulik-Sullivan et al., 2015]. Let S be a set of SNPs and X denote a vector of additively coded genotypes (0, 1 or 2) for a random individual at the SNPs in S. For traits Y_1, Y_2 set

$$\begin{aligned} \beta &= \arg\max_{\alpha \in \mathbb{R}^{|S|}, ||\alpha||=1} \operatorname{cor}(Y_1, X^t\alpha) \\ \gamma &= \arg\max_{\alpha \in \mathbb{R}^{|S|}, ||\alpha||=1} \operatorname{cor}(Y_2, X^t\alpha) \end{aligned} \tag{C.57}$$

where the maximum is taken across the entire population. The genetic correlation between traits across SNPs in S, r_g, is then given by

$$r_g = \frac{\beta^t\gamma}{||\beta||\,||\gamma||} = \sum_{i \in S} \beta_i\gamma_i \tag{C.58}$$

C.5.2 Method 1: control-subgroup 1 vs control-subgroup 2

Expected behaviour

I firstly consider method 1. In this approach, I consider two case-control comparisons:

1. Case subgroup 1 *vs* control group
2. Case subgroup 2 *vs* control group

I denote Z scores derived from GWAS p-values comparing between controls and subgroup 1 by Z_1 and scores between controls and subgroup 2 by Z_2 (figure C.7). An estimated genetic correlation significantly less than 1 (or at least significantly less than estimates from random subgroups) may indicate different causative architectures for the subgroups, in the form of differing relative effect sizes for disease-associated variants.

However, using this method will not distinguish between different disease-causative architectures and genetic differences between subgroups unrelated to the overall phenotype. In terms of the parameters of my three-categories model, method 1 will be liable to reject the null whenever $\tau > 1$, regardless of whether $\sigma_3 > 1$ (that is, regardless of whether subgroup-differentiating SNPs are in general disease-associated). Indeed, for a set value of τ, the negative contribution of SNPs in group 3 to the observed r_g will often be maximised when H_0 holds; that is, $\sigma_3 = 1$.

Consider a SNP in category 3. Under a simple model in which case subgroups are the same size, I denote by μ_c the population MAF of the SNP in controls, and μ_1 and μ_2 the population AF of the same allele in cases. To first order $Z_1 \propto \mu_1 - \mu_c$ and $Z_2 \propto \mu_2 - \mu_c$. Assume $\mu_1 - \mu_2$ is set at some constant $m > 0$. Because $m > 0$, the SNP is associated with at least one of the subgroups, and hence contributes to the genetic correlation. The value of this contribution to the correlation is proportional to $Z_1 Z_2$, which is proportional to $(\mu_1 - \mu_c)(\mu_2 - \mu_c)$.

This is minimised when $\mu_c = \frac{1}{2}(\mu_1 + \mu_2)$. This is exactly the scenario in which the genetic subgroup differences are unrelated to the phenotype as a whole. In other words, dividing the case group on an arbitrary genetically-associated phenotype (ie hair colour, ethnicity, presence of a second unrelated disease) would lead to a lowering of r_g *more* than would a differential disease process with the same heritability (figure C.7).

Simulations

I demonstrated this on the ATD dataset by using the subgroups generated under H_0 as in simulation 1b (see section C.4.2). These subgroups had a true value of τ greater than 1, but $\sigma_3 = 1$ and $\rho = 0$.

For each simulated subgroup, I computed the genetic correlation between the two studies using two methods - LD score regression (LDSC) [Bulik-Sullivan et al., 2015] and genome-wide complex trait analysis (GCTA) [Lee et al., 2012] - and computed the PLR statistic. I also computed genetic correlation and PLR scores for multiple random subgroups of the ATD case group. Significance of the genetic correlation was assessed by either comparing the observed r_g to the values observed in random subgroups (LDSC) or comparing the likelihood of the observed data with an alternative model in which $r_g \equiv 1$.

As expected, r_g estimates using both methods were markedly lower in subgroups with simulated genotypic differences than they were in random subgroups (figure C.8). In the LDSC method, a cutoff of $p < 0.05$ led to rejecting the null in of 45% (SE 2%) of cases, and in GCTA in in 29% (SE 5%) of cases. The PLR method did not reject the null more often than expected, rejecting the null in 4% (SE 1%) of cases.

Application to real data

I also used both LDSC and GCTA to test the hypothesis of differential genetic architecture in GD and HT. The GCTA method was unable to reject the null hypothesis ($p = 0.217$), using a likelihood ratio test against a null model with $r_g = 1$. The LDSC method was able to reject

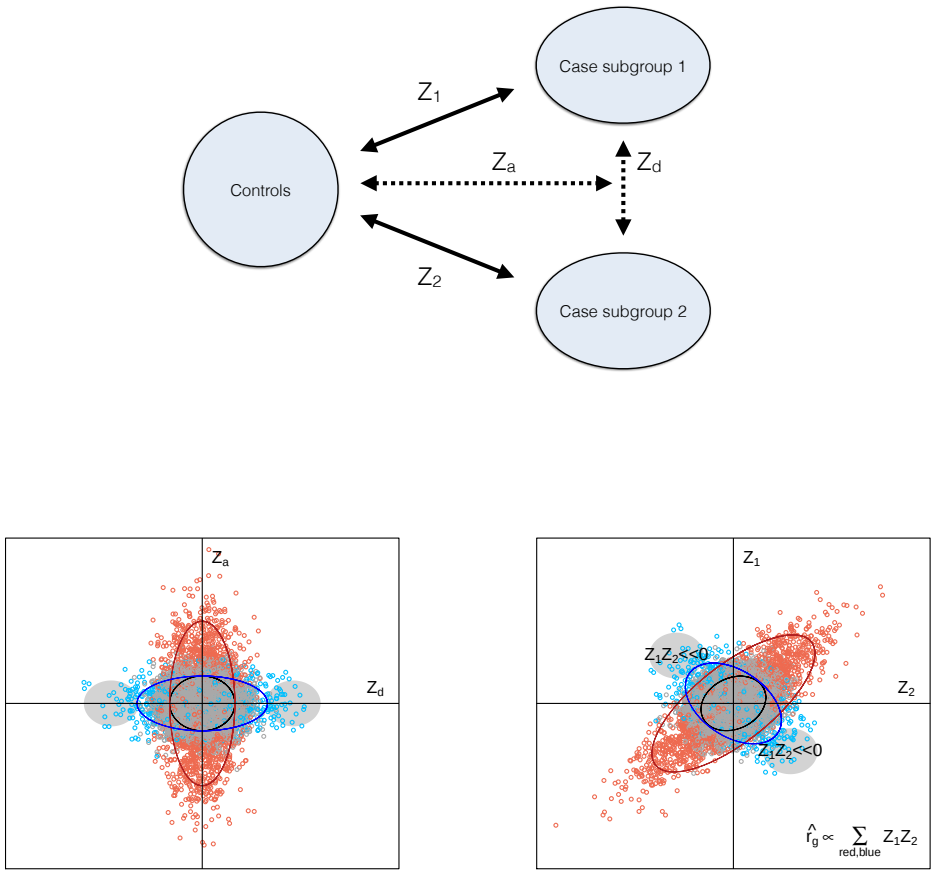


Fig. C.7 One way to test for phenotypic heterogeneity using genetic correlation (r_g) is to estimate r_g for two separate case-control studies; each comparing the control group to one of the disease subgroups, and test whether the estimated r_g is significantly less than 1. I denote by Z_1, Z_2 the sets of Z- scores corresponding to allelic differences between controls and cases of subtype 1 and between controls and cases of subtype 2 respectively (top panel) in contrast to the usual Z_a and Z_d scores. A shortcoming of this method is that r_g is decreased by the presence of SNPs which show allelic differences between subtypes, but are unrelated to the phenotype overall. In this sense, the test $r_g < 1$ is responsive to *any* genetic difference between subtypes - not just those which correspond to differing disease pathology. This scenario would arise if subgroups were defined based on a phenotype with non-zero heritability which was unrelated to the disease; eg, subgroups of T1D defined by hair colouring. The lower two panels demonstrate this scenario. The left panel shows (simulated) Z_a and Z_d scores for a set of SNPs under H_0, where grey corresponds to category 1, red to category 2, and blue to category 3. The right lower panel shows the corresponding sets of Z_1 and Z_2 values. SNPs in the grey circles, and generally SNPs coloured blue, will contribute negatively to the overall genetic correlation, which is asymptotically proportional to the sum of Z_1Z_2 over all SNPs coloured red or blue.

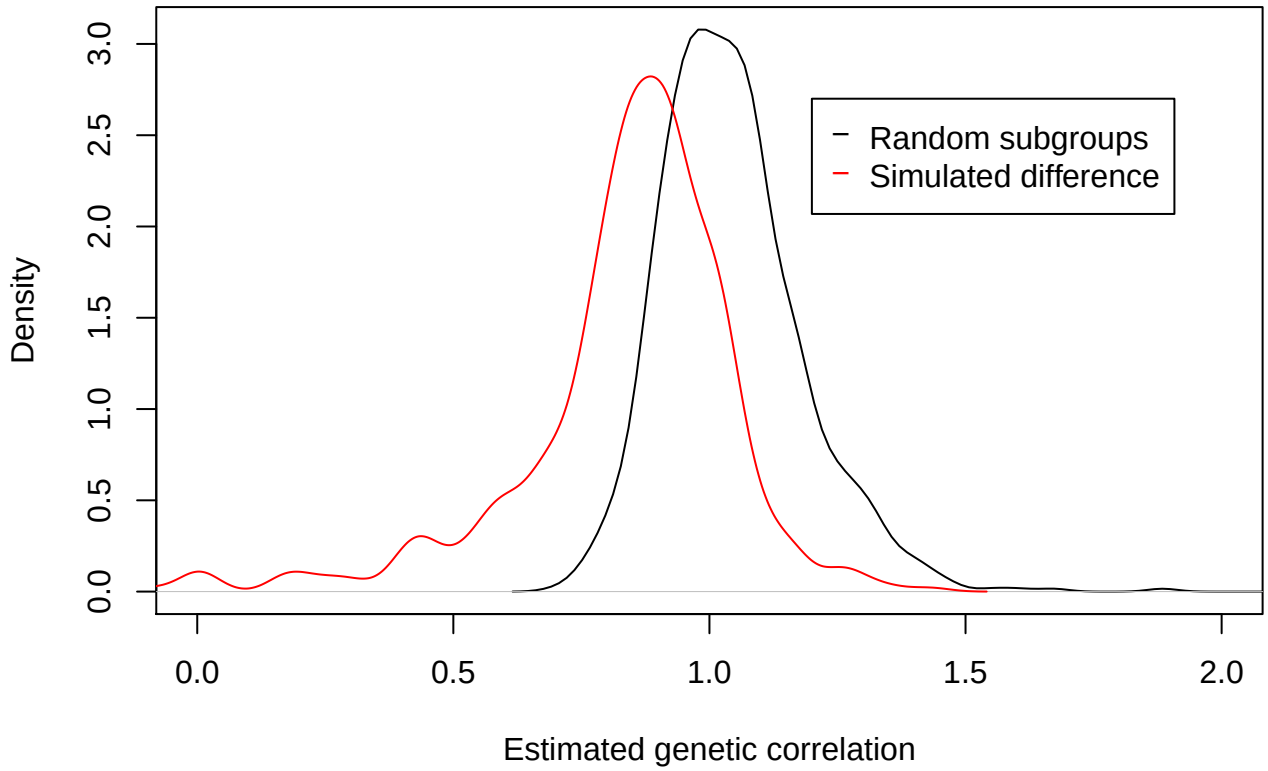


Fig. C.8 Density of estimated r_g (LDSC method) for method 1. Estimates for random subgroups generated under H_0 are shown in black. Estimates for subgroups with a simulated difference ($\tau > 1$) are shown in red. A test based on method 1 would reject H_0 if r_g was significantly less than 1; however, as the plot shows, this would lead to systematic false positives in the scenario where $\tau > 1$. Some estimated values of r_g are greater than 1 due to the way the statistic is estimated under the LDSC method.

the null at $p < 0.05$, though not at the same significance as the PLR (LDSC: $p = 0.012$, PLR $p = 2.2 \times 10^{-15}$). This suggests that the r_g based methods are less powerful than the PLR in this context. This is likely due to the PLR responding to an additional degree of freedom (σ_3) between the null and full models.

C.5.3 Method 2: Z_d (case vs control) vs Z_a (subgroup 1 vs 2)

Expected behaviour, and relation of ρ_g to ρ

In method 2, I consider the two case-control comparisons:

1. Combined case group *vs* control group
2. Case subgroup 1 *vs* case subgroup 2

analogous to my approach in the PLR method, with the two comparisons corresponding to Z_a and Z_d respectively. I estimate r_g between these two traits, and test against the null hypothesis that $r_g = 0$.

The value of r_g relates to the estimated value of ρ_g in the full model. For a set S of disease-associated SNPs with additive (non-epistatic) effects in linkage equilibrium, and a binary trait y, we have

$$\mathrm{cor}(y, X^t\alpha) = \sum_{i \in S} \mathrm{cor}(y, \alpha_i X_i) = \sum_{i \in S} \alpha_i \mathrm{cor}(y, X_i) \tag{C.59}$$

This is maximised when $\alpha_i \propto \mathrm{cor}(y, X_i)$. If $\mu_1(i)$ denotes the AF of SNP i in S amongst the population with $y = 1$, $\mu_0(i)$ the corresponding $\mu_c(i)$ the overall AF of SNP i and p the incidence of the trait in the population (that is, $Pr(y = 1)$), we have

$$\mathrm{cor}(y, X_i) = \sqrt{2p(1-p)} \frac{\mu_1(i) - \mu_0(i)}{\sqrt{\mu_c(i)(1-\mu_c(i))}} \tag{C.60}$$

Given observed allele frequencies $m_1(i)$, $m_0(i)$ at SNP i in a GWAS between traits 1 and 2 with n_1 and n_0 samples respectively, the Z score for significance of that SNP is

$$\begin{aligned} Z(i) &= \frac{m_1(i) - m_0(i)}{SE(m_1(i) - m_0(i))} + O((m_1(i) - m_0(i))^2) \\ &= \frac{m_1(i) - m_0(i)}{\sqrt{\frac{m_1(i)(1-m_1(i))}{n_1} + \frac{m_0(i)(1-m_0(i))}{n_0}}} + O((m_1(i) - m_0(i))^2) \end{aligned} \tag{C.61}$$

so

$$\lim_{\substack{n_1,n_0\to\infty \\ |\mu_1-\mu_0|\to 0}} \left(\frac{1}{n_1}+\frac{1}{n_0}\right)\frac{Z(i)}{\mathrm{cor}(y,X_i)} = \sqrt{p(1-p)} \tag{C.62}$$

Amongst SNPs in LE with small effect sizes ($\mu_1 - \mu_0$ small), expression C.59 is maximised for $\alpha_i \propto \lim_{n_1,n_0\to\infty} Z(i)$. If we denote by Z_{1i}, Z_{2i} the GWAS Z scores for SNP i in phenotypes 1 and 2 respectively in studies with all group sizes $\Theta(n)$, the genetic correlation between the phenotypes is

$$r_g \approx \lim_{n\to\infty} \frac{\sum_{i\in S} Z_{1i}Z_{2i}}{\sqrt{\sum_{i\in S} Z_{1i}^2 \sum_{i\in S} Z_{2i}^2}} \tag{C.63}$$

The sum is over all SNPs S, but the only SNPs with non-vanishing contributions to r_g are those which are associated with both phenotypes. For the two traits in method 2, these SNPs are exactly those which are in my (idealised) category 3 in the full model. Writing C_i as the category of the SNP i we can rewrite the above as

$$r_g \approx \lim_{n\to\infty} \frac{\sum_{i\in S} I(C_i=3) Z_{1i}Z_{2i}}{\sqrt{\sum_{i\in S} I(C_i=3) Z_{1i}^2 \sum_{i\in S} I(C_i=3) Z_{2i}^2}} \tag{C.64}$$

for which an obvious estimator is

$$\hat{r_g} = \frac{\sum_{i\in S} Pr(C_i=3) Z_{1i}Z_{2i}}{\sqrt{\sum_{i\in S} Pr(C_i=3) Z_{1i}^2 \sum_{i\in S} Pr(C_i=3) Z_{2i}^2}} \tag{C.65}$$

If we were to define the full model such that Z_a, Z_d for SNPs in category 3 were distributed as a single bivariate Gaussian distribution with covariance ρ' (as opposed to my current model of two symmetric Gaussians), the updating step for ρ in the E-M algorithm would have a similar form. Indeed, if Θ_{n-1} is the set of estimates for $\{\pi_1, \pi_2, \sigma_2, \sigma_3, \tau, \rho'\}$ after step $n-1$

of the E-M algorithm, the updating steps for ρ', τ, σ_3 are

$$
\begin{aligned}
(\rho')_n &\leftarrow \frac{\sum_{i\in S} Pr(C_i = 3|\Theta_{n-1}) Z_a(i) Z_d(i)}{\sum_{i\in S} Pr(C_i = 3|\Theta_{n-1})} \\
(\sigma_3)_n &\leftarrow \sqrt{\frac{\sum_{i\in S} Pr(C_i = 3|\Theta_{n-1}) Z_a(i)^2}{\sum_{i\in S} Pr(C_i = 3|\Theta_{n-1})}} \\
(\tau)_n &\leftarrow \sqrt{\frac{\sum_{i\in S} Pr(C_i = 3|\Theta_{n-1}) Z_d(i)^2}{\sum_{i\in S} Pr(C_i = 3|\Theta_{n-1})}}
\end{aligned}
\tag{C.66}
$$

and hence when the E-M algorithm converges, $\rho'/(\sigma_3\tau)$ is an estimator for r_g. Testing $r_g \neq 0$ in this scenario is broadly equivalent to testing whether $\rho' \neq 0$ in the adapted full model.

When developing the PLR method, I chose not to use this simpler model, opting for a more complex two-Gaussian distribution of (Z_a, Z_d) for SNPs in category 3. There were several reasons for this choice. Importantly, $\rho' \neq 0$ implies $\rho > 0$, so the test $r_g \neq 0$ tests a more specific proposition than the PLR.

Testing for $\rho' \neq 0$ or $r_g \neq 0$ is weakened when Z_a and Z_d are correlated at some group of SNPs and anticorrelated at others. I note that this simultaneous correlation and anticorrelation is likely in many biological scenarios. Given two disease subgroups 1 and 2, deleterious variants associated only with subgroup 1 will have correlated Z_a, Z_d values, whereas deleterious variants associated only with subgroup 2 will have anticorrelated Z_a and Z_d.

In addition, the presence of between-subgroup heterogeneity, as characterised by the presence of SNPs with simultaneously high $|Z_d|$ and $|Z_a|$ values, does not require that Z_a and Z_d have to be correlated or anticorrelated at all. The presence of a set of SNPs whose marginal variances of Z_a and Z_d are simultaneously significantly larger than 1 is sufficient evidence for heterogeneity of disease basis. This was the impetus for including the additional parameter σ_3 in the full model.

Uncorrelated Z_a and Z_d may well occur in situations where the main sources of variation between the subgroups are only weakly associated with the overall phenotype, while less associated variants are strongly associated. This would be expected to occur in situations where the subtypes have known genetic differences. If, for example, a subgrouping phenotype was based on visual acuity in the phenotype of symptomatic Type 2 diabetes, variants associated with general macular degeneration would have large $|Z_d|$ scores with low $|Z_a|$

scores, while variants associated with microvascular glucose sensitivity would have larger $|Z_a|$ scores and smaller (but still overdispersed) $|Z_d|$ scores.

The behaviours of r_g/ρ', ρ, τ and σ in various scenarios are summarised in appendix D.1, table D.1. Overall, I consider that while ρ_g is a useful statistic, it does not capture the variety of forms that disease heterogeneity can take.

Simulations

I tested the ability of GCTA to reject the null hypothesis $r_g = 0$ on simulated data. I simulated genotypes for 4000 controls and 2000 cases in each of two subgroups at 10000 SNPs in linkage equilibrium. Genotypes were simulated in such a way that Z_a and Z_d scores would have the distributions

$$\begin{aligned}
\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) && \text{at 7000 SNPs } (\pi_1 = 0.7) \\
\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}\right) && \text{at 2000 SNPs } (\pi_2 = 0.1) \\
\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & \rho \\ \rho & 4 \end{pmatrix}\right) && \text{at } \xi * 1000 \text{ SNPs } (\pi_3 = 0.2) \\
\begin{pmatrix} Z_d \\ Z_a \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & -\rho \\ -\rho & 4 \end{pmatrix}\right) && \text{at } (1-\xi) * 1000 \text{ SNPs } (\pi_3 = 0.2)
\end{aligned}$$

The value ξ represents the degree to which Z_a, Z_d scores can show both correlation and anticorrelation, and ρ represents the extent of the correlation/anticorrelation. I ran simulations at $\rho = 0$ and for $\rho \in \{0, 0.5, 1, 2\}$ for $\xi = 0$ (no anticorrelation), $\xi = 0.2$ (mostly correlation, some anticorrelation) and $\xi = 0.5$ (equal correlation and anticorrelation). The large value of π_3 was to ensure that both PLR and GCTA should be well-powered to reject the null hypothesis where able, but not so well-powered as to be incomparable.

I estimated r_g using the GCTA method [Lee et al., 2012]. Significance was assessed using the provided likelihood-ratio test comparing the fitted model with a null model in which $r_g = 0$.

I did not test LDSC in this scenario, as it estimates r_g based on phenomena arising from the LD matrix, and simulation would entail setting an inherent effect size for these phenomena through specifying an LD matrix. Since the shortcomings I identify are with the use of r_g itself, rather than the method used to simulate it, I considered this reasonable.

ρ	ξ	**GCTA**	**PLR**
0	0	0.09 (0.002)	1 (-)
0	0.2	0.12 (0.002)	1 (-)
0	0.5	0.06 (0.002)	1 (-)
0.5	0	0.55 (0.006)	1 (-)
0.5	0.2	0.13 (0.004)	1 (-)
0.5	0.5	0.06 (0.001)	1 (-)
1	0	0.96 (0.002)	1 (-)
1	0.2	0.59 (0.005)	1 (-)
1	0.5	0.07 (0.002)	1 (-)
2	0	1 (-)	1 (-)
2	0.2	1 (-)	1 (-)
2	0.5	0.04 (0.001)	1 (-)

Table C.3 Power of tests to reject the null hypothesis at $\alpha = 0.05$ in simulated data. Brackets show standard error. Value ρ is the degree of correlation/anticorrelation between Z_d and Z_a. Value ξ is the degree of split between correlation and anticorrelation; $\xi = 0$ corresponds to correlation only, $\xi = 0.2$ to mostly correlation with some anticorrelation, and $\xi = 0.5$ to a half/half mix. Testing for subgroup heterogeneity using GCTA is adequately powerful when correlation ρ is present, but declines markedly when both correlation and anticorrelation are present, and is effectively zero when $p = 0.5$ or $\rho = 0$. The PLR-based test was able to reject H_0 universally in all cases.

As expected, the test based on $r_g = 0$ was not able to reject the null hypothesis when $\rho = 0$ or $\xi = 0.5$, and power was markedly reduced when some anticorrelation was present, at $\xi = 0.2$ (figure C.9, table C.3). While the test was able to systematically reject the null hypothesis when $\xi \in \{0, 0.2\}$, $\rho > 0$, the power was universally lower than that of the PLR test (table C.3). This was likely due to information gained from the additional degree of freedom (σ_3) between the full and null models in the PLR test. I did not simulate any scenarios where $\sigma_3 = 1$, as this would imply that SNPs in category 3 were not systematically associated with the subgrouping phenotype, and hence correlation with Z_a would be spurious.

Application to real data

Finally, I assessed whether I could reject H_0 by testing against $r_g = 0$ on my ATD dataset (MHC removed), with subtypes GD and HT. I used both the LDSC and GCTA methods to do this. While both were able to reject the null hypothesis (LDSC: $r_g = -0.579, p = 0.04$, from known null distribution of ρ_g; GCTA: $r_g = -0.580, p = 1 \times 10^{-3}$ from likelihood-ratio test) neither could do so as confidently as the PLR test ($p = 2.2 \times 10^{-15}$).

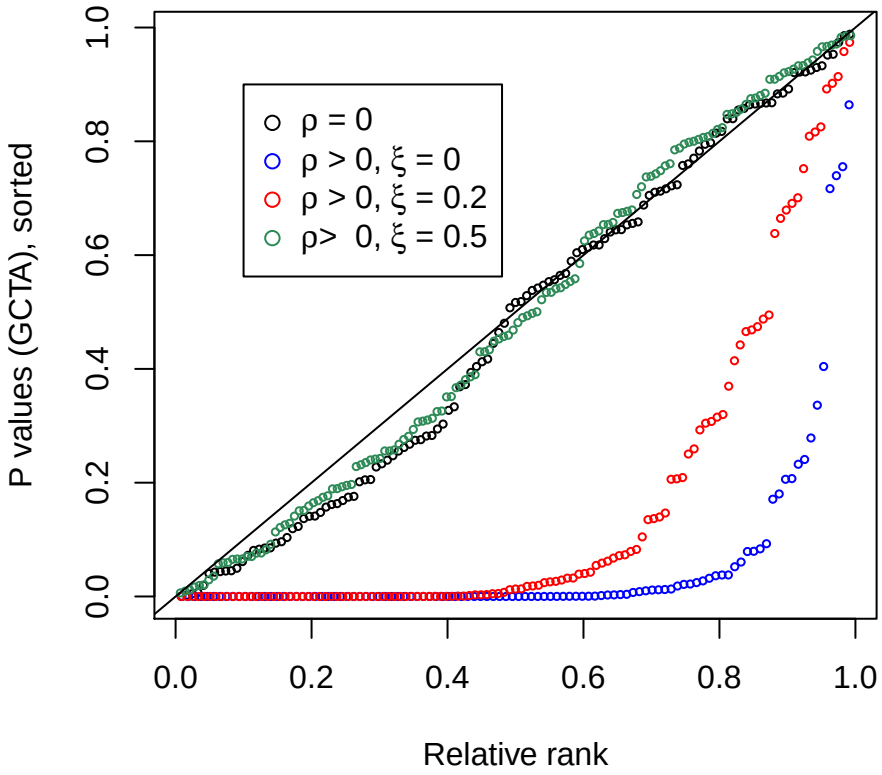


Fig. C.9 Sorted p values from test of null hypothesis $r_g = 0$ under simulations in which $\rho \in \{0, 0.5, 1, 2\}$ and $\xi \in \{0, 0.2, 0.5\}$. In all simulations, H_0 is false (with $\sigma_3 > 0$). GCTA is able to reject the null hypothesis only if $\rho > 0$ and $p \neq 0.5$, and power is reduced (ie, p-values are higher) if $p = 0.2$ compared to $p = 0$. If $\rho = 0$ or $\xi = 0.5$, the p-values show effectively no deviation from $U(0,1)$. Thus a test based on rejecting $\rho_g = 0$ is not suitable for my purposes.

Unfortunately I was unable to obtain meaningful estimates of r_g in comparisons relating to type 1 diabetes subgroups; the standard error of the estimate was such that I was never able to reject the null hypothesis $r_g = 1$ (method 1) or $r_g = 0$ (method 2), with all adjusted p-values greater than 0.2. Given this, I do not consider the estimated r_g values to be informative in this case.

Our proposed test is complex, and parametrises disease heterogeneity using several variables (namely π_3, σ_3, τ and ρ) rather than providing a single metric. I consider this complexity to be necessary; heterogeneity in a phenotype can arise in many ways and the heterogeneous genetic architecture can take many forms. A test specifically to detect SNPs with large, genome-wide significant effect sizes in one disease subgroup but not the other may miss heterogeneity characterised by subtle effect size differences across many SNPs with small effects. My method can ideally detect heterogeneity in a general sense in multiple situations, and give insight into the architecture in the form of the fitted parameters.

C.6 Other

C.6.1 Alternative test statistics for retrospective single-SNP analysis

I propose four summary statistics for testing the degree to which single SNPs have differential effect sizes in disease subgroups. The fourth of these, the Bayesian conditional false discovery rate (cFDR) is discussed in the methods section of the main text. The three alternative statistics (which I term X_1, X_2, X_3) test against slightly different null hypotheses. The alternative statistic for testing the overall H_0 described in section 5.5.2 can also be adapted to retrospective single-SNP analysis, although it is difficult to interpret its meaning for individual SNPs.

The first, X_1, is the posterior probability of membership of the third category of SNPs under the full model; that is, for a SNP of interest with Z scores z_a, z_d and given fitted parameters $\Theta_1 = \{\pi_1, \pi_2, \pi_3, \sigma_2, \sigma_3, \tau, \rho\}$:

$$\begin{aligned} X_1 &= Pr(\text{SNP} \in \text{category } 3 | \Theta_1) \\ &= \frac{\frac{1}{2}\pi_3 \left(N_{\mathbf{0}, \begin{pmatrix} \tau^2 & \rho \\ \rho & \sigma_3^2 \end{pmatrix}}(z_a, z_d) + N_{\mathbf{0}, \begin{pmatrix} \tau^2 & -\rho \\ -\rho & \sigma_3^2 \end{pmatrix}}(z_a, z_d) \right)}{PDF_{\Theta_1}(z_a, z_d)} \end{aligned} \tag{C.67}$$

This test statistic has the advantage of straightforward FDR control against the null hypothesis $H_0 = \{\text{SNP} \in \text{category } 1/2|\Theta_1\}$, assuming the validity of Θ_1. It also reflects the overall shape of the distribution. A disadvantage is the dependence on the model implied by Θ_1; in circumstances where $\sigma_3 >> \sigma_2$, the test statistic X_1 will be high for high values of $|Z_a|$ even when $|Z_d|$ is low (appendix D.2, figures D.8). This is a particular problem if tested regions include very strong associations; for example, the MHC region in autoimmune phenotypes.

Our second statistic, X_2, is the difference in pseudo-log likelihood of a given SNP under the full and null models; that is, given fitted parameters Θ_1 under H_1 and Θ_0 under H_0

$$X_2 = \log\{PL(z_a, z_d|\Theta_1)\} - \log\{PL(z_a, z_d|\Theta_0)\} \tag{C.68}$$

This has the advantage that high values of X_2 directly identify the SNPs contributing to a higher pseudo-likelihood ratio. A disadvantage is the sensitivity to the behaviour of the fitted parameters under H_0, which may be variable (see chapter 5, section 5.2.3 and table 5.2), and absence of direct FDR control. Because X_1 and X_2 tend to highlight uninteresting SNPs in differing circumstances, I found a combination of both to be useful to find SNPs which are 'unusual' (high X_1) and contribute to the PLR (high X_2).

The third test statistic is defined as $X_3 = z_a^{\alpha} z_d^{1-\alpha}$, $\alpha \in (0,1)$. I chose this test statistic as I am broadly searching for evidence of correlation between Z_a and Z_d, and SNPs contribute to measures of correlation principally through the value of $Z_a Z_d$. This test statistic identifies SNPs with concurrently high Z_a and Z_d in an obvious way, so is of most use when SNPs which differentiate subgroups are not of interest unless they are also associated with the overall phenotype.

The value of α is set in order to prioritise SNPs with high Z_d over those with high Z_a; for instance, with $\alpha = 0.5$ will give equal weight to a SNP with $Z_a = 10$, $Z_d = 1$ and a SNP with $Z_a = 1$, $Z_d = 10$, but in general the second SNP will be of far greater interest. To determine the best value of α, I consider how much I may expect Z_a and Z_d to deviate from 0, using both the full and null models.

I set τ' as the largest value of τ across both models, and σ' as the largest of σ_2 (null model) and σ_2, σ_3 (full model). Given fitted values τ', σ', I suggest the value

$$\alpha = \frac{\log(\sigma')}{\log(\tau') + \log(\sigma')} \tag{C.69}$$

so that the statistic X_3 has the same value at the points $(1, \tau')$ and $(\sigma', 1)$. The rationale for this is that SNPs which have the true underlying distributions $N_{\mathbf{0},\left(\begin{smallmatrix}\tau'^2 & 0\\ 0 & 1\end{smallmatrix}\right)}$ or $N_{\mathbf{0},\left(\begin{smallmatrix}1 & 0\\ 0 & \sigma'^2\end{smallmatrix}\right)}$ are

uninteresting; I seek deviance from both of these distributions. A hypothesis test for X_3 can then be computed, using the appropriate values of $\pi_{(0,1,2)}$.

Contour plots of the test statistics for several datasets are shown in appendix D.2, figures D.8, D.9.

C.6.2 Independence of PLR distribution on subgroup sizes

PLR and cPLR values for randomly chosen subgroups are all derived from data with the same Z_a values, with the distribution of Z_d expected to be $N(0,1)$ and independent of Z_a regardless of the relative sizes of random subgroups. Therefore I expect that the asymptotic distribution (chapter 5, equation 5.2 does not depend on relative subgroup size. An important consequence of this is that if several subgroupings of a phenotype are being simultaneously assessed, the empirical distribution of cPLR need only be calculated once.

I demonstrate this assertion by simulation. Using the autoimmune thyroid disease dataset, I simulated random subgroups from the combined case group (GH+HT) for a range of relative sizes, repeating the simulation 1000 times for each subgroup size. Figure C.10 shows the observed distributions of PLR and cPLR as compared to the overall distribution. These plots are consistent with independence of empirical PLR and cPLR distributions on subgroup size.

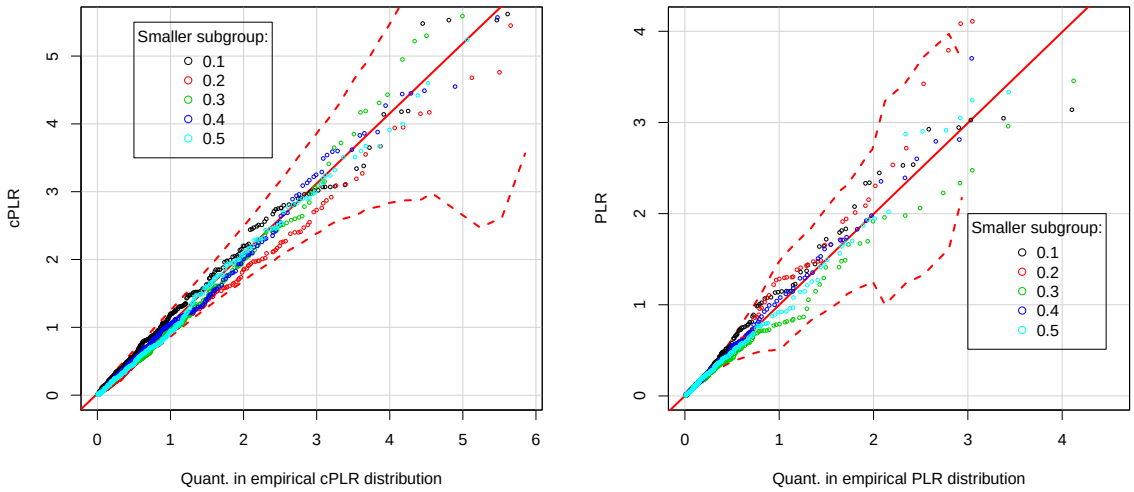


Fig. C.10 Distributions of PLR and cPLR for various relative sizes of subgroups. Simulations are on ATD data. Legend shows the proportion of cases in the smaller subgroup. Leftmost plot shows distribution of observed cPLR, rightmost distribution of PLR. Red dotted lines show empirical 99% confidence limits. Distributions are similar for all relative subgroup sizes.

C.6.3 Number of simulations necessary to fit null distribution

I assessed the number of simulated random subgroups required to estimate the parameters γ, κ of the null distribution of the cPLR. I took bootstrap samples of various sizes from my list of simulated random subgroups ($\tau = 1$) of the ATD data. For each sample, I computed the fitted values of γ and κ and the observed p-values associated with observed PLR values of 2, 3, 5, and 10, i.e. expected p values 0.08, 0.03, 0.004 and 1.5×10^{-6} respectively (figure C.11)

This suggests that 1000 simulations is generally adequate, and it is difficult to improve accuracy markedly past this point. For this number of simulations, 95% of computed values for κ, γ, $Pr(PLR > 2|\kappa,\gamma)$ and $Pr(PLR > 5|\kappa,\gamma)$ were in $[0.44, 0.56]$, $[0.46, 0.72]$, $[0.069, 0.97]$ and $[0.0021, 0.0057]$ respectively. As expected, consistency of p-value estimates is poorer for lower p-values, as these correspond to greater extrapolations of the distribution.

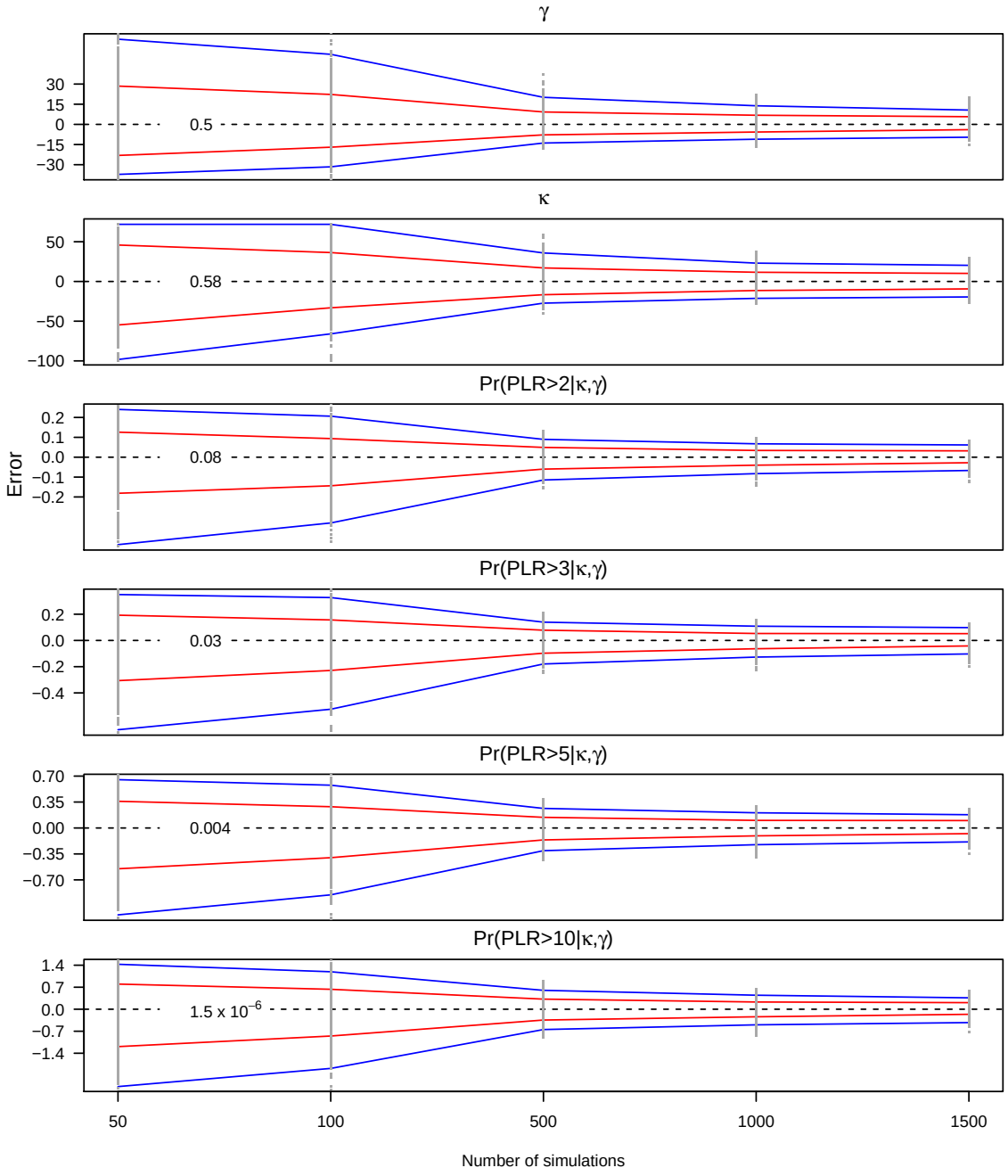


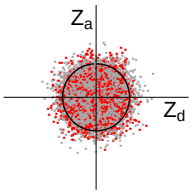
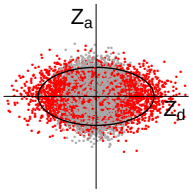
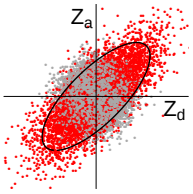
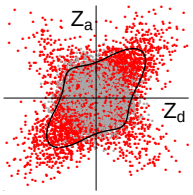
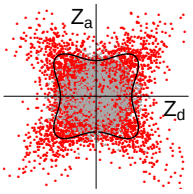
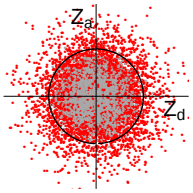
Fig. C.11 Distributions of estimated parameters γ and κ and various corresponding p-values, using various sizes of simulated random subgroups. Blue lines show quantiles of observed distribution corresponding to $\pm 2\sigma$; red lines show quantiles corresponding to $\pm\sigma$. Errors in γ and κ are shown as percentage errors as compared to median. Errors in p-values are shown as $\log_{10}$ fold changes from median. Values of the median value of each variable are shown. Observed values are shown as grey dots.

Appendix D

Supplementary tables and figures for chapter 5

D.1 Supplementary tables

Please see following page

Form	$r_g^{(1)}$	$r_g^{(2)}$	ρ	τ	σ_3	Phenomenon
	1	0	0	> 1	1	H_0: $Z_d, Z_a \sim N(0, I_2)$; all-environmental cause for subgroup phenotype
	$\ll 1$	0	0	> 1	1	H_0: Z_d, Z_a independent; subgrouping phenotype independent of main phenotype;
	$1/ < 1$	$\gg 0$	$\gg 0$	> 1	> 1	H_1: Z_d, Z_a correlated; eg. same pathways; different heritability (age-of-onset)
	< 1	> 0	$\gg 0$	> 1	> 1	H_1: Z_d, Z_a mostly correlated, some anticorrelation; eg. most variants associated with subgroup 1, some with subgroup 2
	< 1	0	$\gg 0$	> 1	> 1	H_1: Z_d, Z_a both correlated and anticorrelated; eg. variants either associated only with subgroup 1 or only with subgroup 2
	< 1	0	0	> 1	> 1	H_1: $\text{var}(Z_d) > 1$ and $\text{var}(Z_a) > 1$ but not correlated; general shared genetic architecture between subgrouping phenotype and main phenotype, effect sizes independent

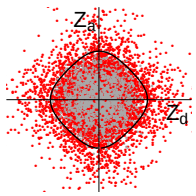
	< 1	0	0	> 1	> 1	H_1: shared genetic architecture between subgrouping phenotype and main phenotype, effect sizes dependent but not correlated or anticorrelated

Table D.1 Heterogeneity between case subgroups may arise in multiple ways, some of which are illustrated here. Plots show the distribution of Z_d and Z_a for SNPs in category 3 (those which differentiate subgroups). Column $r_g^{(1)}$ corresponds to genetic correlation in method 1 (between Z scores for control vs subgroup 1 and control vs subgroup 2), and column $r_g^{(2)}$ to genetic correlation in method 2 (between Z_a and Z_d); see appendix C, section C.5. SNPs in category 1 (not differentiating cases/controls and not differentiating subgroups) are shown in grey for reference, and SNPs in category 2 are omitted. In the first two rows, the pathology leading to heterogeneity is genetically independent of the pathology leading to the main phenotype; the main null hypothesis. The test $r_g^{(1)} < 1$ will reject H_0 for the scenario in row 2, as well as other scenarios. The test $r_g^{(2)} \neq 0$ rejects H_0 for the scenario in row 3, but is weakened in the scenario in row 4 due to the anticorrelation, and will not be able to reject H_0 for rows 5-7. Since ρ detects correlation and anticorrelation simultaneously, it will additionally reject H_0 for row 4 and will not be weakened in row 3. However, it is necessary to test for $\sigma_3 > 1$ to reject H_0 for rows 5 and 6.

	Model	π_1	π_2	π_3	σ_2	σ_3	τ	ρ	p-value
TPO-Ab	Full	0.511	0.487	2.407×10^{-3}	0.994	6.545	1.552	0.991	$< 1 \times 10^{-20}$
	Null	0.987	2.333×10^{-3}	0.011	6.634	-	1.308	-	
TPO-Ab	Full	0.997	2.898×10^{-4}	3.031×10^{-3}	4.698	2.291	1.497	0.338	1.5×10^{-4}
no MHC	Null	0.989	1.882×10^{-3}	9.087×10^{-3}	3.11	-	1.318	-	
GAD-Ab	Full	0.995	3.557×10^{-3}	1.057×10^{-3}	2.832	8.866	2.295	5.484	$< 1 \times 10^{-20}$
	Null	0.997	2.328×10^{-3}	3.002×10^{-4}	6.639	-	2.153	-	
GAD-Ab	Full	0.997	2.9×10^{-3}	3.434×10^{-4}	2.279	4.531	1.055	3.424	0.002
no MHC	Null	0.792	1.883×10^{-3}	0.206	3.111	-	0.997	-	
IA2-Ab	Full	0.995	3.275×10^{-3}	1.244×10^{-3}	2.804	8.291	3.027	1.575	$< 1 \times 10^{-20}$
	Null	0.997	2.287×10^{-3}	3.805×10^{-4}	6.674	-	3.852	-	
IA2-Ab	Full	0.998	1.362×10^{-3}	7.904×10^{-4}	3.318	2.212	2.145	0	0.008
no MHC	Null	0.998	1.88×10^{-3}	2.073×10^{-4}	3.112	-	2.889	-	
PCA-Ab	Full	0.997	2.336×10^{-3}	3.413×10^{-4}	6.631	0.37	2.097	0.422	> 0.5
	Null	0.998	2.335×10^{-3}	1.276×10^{-4}	6.632	-	2.54	-	
PCA-Ab	Full	0.997	2.759×10^{-3}	1.303×10^{-4}	2.508	5.58	2.256	0	> 0.5
no MHC	Null	0.998	1.884×10^{-3}	1.384×10^{-4}	3.111	-	2.5	-	

Table D.2 Parameters of models fitted to T1D autoantibody positivity data. With MHC retained (co-ordinates 25-38 Mb, GChR build 37) all full models fit better than null models with the exception of those fitted to PCA-Ab positivity. With MHC removed, effect sizes were lower, but the null hypothesis could be rejected for TPOA-Ab positivity, with weaker evidence for rejecting the null hypothesis for GAD-Ab and IA2-Ab. In most cases, there was evidence of SNPs differentiating subgroups (typically, fitted $\tau > 1$). There were generally a small number of SNPs which strongly differentiated cases and controls (a small value of π_2, π_3 corresponding to the larger value of σ_2, σ_3). P-values were computed against the null distribution of cPLR for random subgroups, which showed good agreement with the asymptotic mixture-χ^2 distribution (see appendix D.2, figure D.5c. P-values shown are unadjusted for multiple testing.

Age	Full	0.898	0.099	2.4×10^{-3}	0.96	6.558	1.601	3.644	4.9×10^{-37}
	Null	0.885	2.338×10^{-3}	0.113	6.631	-	0.945	-	
Age	Full	0.997	1.881×10^{-4}	3.035×10^{-3}	5.257	2.372	1.159	1.315	0.007
no MHC	Null	0.782	1.891×10^{-3}	0.216	3.107	-	0.97	-	

Table D.3 Parameters of models fitted to age at diagnosis in T1D, considered as a parameter rather than defining subgroups. The full model fit significantly better than the null model when the MHC region was included or excluded. Plotted Z_a and Z_d scores are shown in appendix D.2 figure D.7. The fitted models show evidence of SNPs associated with age at diagnosis (fitted $\tau > 1$). P-values were computed against the null distribution of cPLR for random subgroups, which showed good agreement with the asymptotic mixture-χ^2 distribution (see appendix D.2, figure D.5c).

SNP details				Z scores	
SNP	Chr	Pos	Gene	$\lvert Z_d \rvert (p)$	$\lvert Z_a \rvert$
rs12045559	1	113708908	*PTPN22*	2.892 (3.8×10^{-3})	3.217
rs415024	5	9445358		3.051 (2.3×10^{-3})	2.941
rs1010599	5	35944231	*IL7R*	3.367 (7.6×10^{-4})	2.881
rs4024109	5	35955375	*IL7R*	3.307 (9.4×10^{-4})	2.792
rs17085170	5	95198087		4.291 (1.8×10^{-5})	2.365
rs3114834	7	109192112		2.649 (8.1×10^{-3})	3.787
rs12549890	8	21045174		3.535 (4.1×10^{-4})	2.459
rs16874205	8	107271324		4.428 (9.5×10^{-6})	3.565
rs4076319	10	85129122		4.124 (3.7×10^{-5})	2.165
rs10736277	10	121705898		3.415 (6.4×10^{-4})	3.411
rs7912574	10	121717404		3.265 (1.1×10^{-3})	2.957
rs2065660	10	121754185		3.557 (3.8×10^{-4})	3.031
rs6578252	11	2226817	*INS*	3.481 (5×10^{-4})	2.576
rs705698	12	54670954	*IKZF4*	5.058 (4.2×10^{-7})	2.016
rs705702	12	54676903	*IKZF4*	5.135 (2.8×10^{-7})	2.086
rs2292239	12	54768447	*IKZF4*	5.651 (1.6×10^{-8})	3.278
rs4766443	12	109864518		3.372 (7.5×10^{-4})	3.644
rs10774613	12	110008885	*SH2B3*	3.929 (8.5×10^{-5})	2.614
rs1265566	12	110179096	*SH2B3*	3.612 (3×10^{-4})	3.975
rs17696736	12	110949538	*SH2B3*	3.867 (1.1×10^{-4})	6.409
rs16961362	15	33731898		3.799 (1.5×10^{-4})	3.23
rs1711029	15	51491702		5.178 (2.2×10^{-7})	4.201
rs12924729	16	11095284	*DEXI*	3.741 (1.8×10^{-4})	3.784
rs1942707	18	60768535		4.279 (1.9×10^{-5})	1.642

SNP details		Values (rank)			
SNP	Chr	X_1	X_2	X_3	X_4
rs12045559	1	0.332	0.383	3.003	0.061 (15)
rs415024	5	0.344	0.394	3.012	0.083 (20)
rs1010599	5	0.538	0.72 (15)	3.186 (16)	0.078 (18)
rs4024109	5	0.456	0.561 (18)	3.115 (19)	0.107
rs17085170	5	0.832	1.353 (9)	3.477 (10)	0.022 (12)

rs3114834	7	0.308	0.351	3.005	0.072 (17)
rs12549890	8	0.449	0.519 (19)	3.11 (20)	0.141
rs16874205	8	0.994	4.48 (3)	4.102 (4)	6.669×10^{-4} (3)
rs4076319	10	0.656	0.773 (14)	3.285 (15)	0.066 (16)
rs10736277	10	0.775	1.434 (8)	3.413 (12)	0.021 (11)
rs7912574	10	0.499	0.648 (16)	3.153 (17)	0.084
rs2065660	10	0.73	1.23 (11)	3.361 (14)	0.036 (13)
rs6578252	11	0.473	0.572 (17)	3.13 (18)	0.113
rs705698	12	0.934	0.871 (13)	3.656 (8)	1.241×10^{-3} (6)
rs705702	12	0.956	1.067 (12)	3.737 (6)	8.403×10^{-4} (4)
rs2292239	12	1	4.991 (2)	4.663 (2)	8.184×10^{-6} (1)
rs4766443	12	0.813	1.623 (7)	3.466 (11)	0.016 (10)
rs10774613	12	0.769	1.279 (10)	3.403 (13)	0.055 (14)
rs1265566	12	0.942	2.764 (5)	3.736 (7)	6.384×10^{-3} (8)
rs17696736	12	0.237	0.256	4.622 (3)	9.922×10^{-4} (5)
rs16961362	15	0.896	2.123 (6)	3.588 (9)	0.011 (9)
rs1711029	15	1	7.503 (1)	4.809 (1)	8.759×10^{-6} (2)
rs12924729	16	0.952	2.92 (4)	3.756 (5)	3.989×10^{-3} (7)
rs1942707	18	0.411	0.108	3.052	0.083 (19)

SNP details		Summary statistics	
SNP	Chr	p-val (X_3)	FDR (X_4)
rs12045559	1	5.593×10^{-5}	0.169
rs415024	5	5.415×10^{-5}	0.247
rs1010599	5	2.092×10^{-5}	0.23
rs4024109	5	3.072×10^{-5}	0.357
rs17085170	5	4.695×10^{-6}	0.046
rs3114834	7	5.504×10^{-5}	0.213
rs12549890	8	3.154×10^{-5}	0.48
rs16874205	8	2.768×10^{-7}	3.612×10^{-3}
rs4076319	10	1.269×10^{-5}	0.186
rs10736277	10	6.581×10^{-6}	0.045
rs7912574	10	2.528×10^{-5}	0.25
rs2065660	10	8.529×10^{-6}	0.091
rs6578252	11	2.87×10^{-5}	0.374

rs705698	12	2.016×10^{-6}	6.997×10^{-3}
rs705702	12	1.371×10^{-6}	4.78×10^{-3}
rs2292239	12	2.35×10^{-8}	2.586×10^{-5}
rs4766443	12	5.058×10^{-6}	0.069
rs10774613	12	6.902×10^{-6}	0.152
rs1265566	12	1.373×10^{-6}	0.027
rs17696736	12	2.823×10^{-8}	5.809×10^{-3}
rs16961362	15	2.769×10^{-6}	0.052
rs1711029	15	1.187×10^{-8}	2.748×10^{-5}
rs12924729	16	1.278×10^{-6}	0.015
rs1942707	18	4.376×10^{-5}	0.247

Table D.6 Top 20 SNPs differentiating T1D and RA (MHC removed), considered as subgroups of a general autoimmune phenotype, for each of four summary statistics. Positions are in NCBI build 36. Because of the large number of SNPs with evidence for differentiating the subgroups, only SNPs with non-zero weights after applying the LDAK procedure are included in this table. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0' \big| |Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores	
SNP	Chr	Pos	Gene	$\lvert Z_d\rvert(p)$	$\lvert Z_a\rvert$
rs17013326	1	113801358	*PTPN22*	3.007 (2.6×10^{-3})	4.509
rs1230666	1	113885452	*PTPN22*	4.327 (1.5×10^{-5})	6.265
rs6679677	1	114015850	*PTPN22*	6.52 (7×10^{-11})	7.437
rs6661817	1	114159076	*PTPN22*	3.372 (7.5×10^{-4})	3.26
rs3811019	1	114183625	*PTPN22*	3.353 (8×10^{-4})	3.609
rs12061474	1	201120971	*PIK3C2B*	3.252 (1.1×10^{-3})	3.153
rs903228	2	53603700		0.077 (0.94)	5.085
rs7666328	4	116140909		0.425 (0.67)	5.593
rs2544677	5	86435018		3.272 (1.1×10^{-3})	3.661
rs2112168	5	86440646		3.199 (1.4×10^{-3})	3.335
rs7917983	10	114722872	*TCF7L2*	4.357 (1.3×10^{-5})	3.303
rs7901275	10	114722896	*TCF7L2*	4.331 (1.5×10^{-5})	3.287
rs7901695	10	114744078	*TCF7L2*	7.691 (1.5×10^{-14})	2.215
rs12243326	10	114778805	*TCF7L2*	6.645 (3×10^{-11})	1.889
rs3741939	12	3517792		4.885 (1×10^{-6})	0.473
rs705698	12	54670954	*IKZF4*	4.494 (7×10^{-6})	2.569
rs705702	12	54676903	*IKZF4*	4.554 (5.3×10^{-6})	2.624
rs2292239	12	54768447	*IKZF4*	5.026 (5×10^{-7})	3.78
rs4766443	12	109864518	*SH2B3*	3.44 (5.8×10^{-4})	3.025
rs10774613	12	110008885	*SH2B3*	4.415 (1×10^{-5})	2.223
rs1265566	12	110179096	*SH2B3*	3.398 (6.8×10^{-4})	3.276
rs17696736	12	110949538	*SH2B3*	4.981 (6.3×10^{-7})	4.788
rs12924729	16	11095284	*SH2B3*	4.236 (2.3×10^{-5})	3.563
rs7193144	16	52368187	*FTO*	4.493 (7×10^{-6})	2.139
rs8050136	16	52373776	*FTO*	4.442 (8.9×10^{-6})	2.127
rs9926289	16	52378004	*FTO*	4.19 (2.8×10^{-5})	1.985
rs2542151	18	12769947	*PTPN2*	5.278 (1.3×10^{-7})	2.866

SNP details		Values (rank)			
SNP	Chr	X_1	X_2	X_3	X_4
rs17013326	1	0.957	3.169 (7)	3.741 (9)	3.431×10^{-3} (13)
rs1230666	1	1	14.651 (2)	5.283 (2)	1.512×10^{-5} (6)

rs6679677	1	1	22.963 (1)	6.999 (1)	7.052×10^{-11} (1)
rs6661817	1	0.573	0.801	3.311 (17)	0.021
rs3811019	1	0.768	1.413 (14)	3.489 (11)	9.448×10^{-3} (17)
rs12061474	1	0.422	0.511	3.198 (20)	0.036
rs903228	2	0.726	1.425 (13)	0.738	≥ 0.5
rs7666328	4	0.954	3.307 (6)	1.706	≥ 0.5
rs2544677	5	0.753	1.36 (15)	3.476 (12)	9.796×10^{-3} (18)
rs2112168	5	0.503	0.664	3.272 (18)	0.03
rs7917983	10	0.973	2.988 (8)	3.752 (8)	8.059×10^{-4} (9)
rs7901275	10	0.969	2.895 (9)	3.732 (10)	8.347×10^{-4} (10)
rs7901695	10	1	1 (16)	3.93 (5)	1.566×10^{-10} (2)
rs12243326	10	1	0.447 (20)	3.372 (14)	7.82×10^{-8} (3)
rs3741939	12	0.713	0	1.386	0.016 (19)
rs705698	12	0.907	1.492 (12)	3.324 (16)	2.223×10^{-3} (12)
rs705702	12	0.932	1.673 (11)	3.383 (13)	1.886×10^{-3} (11)
rs2292239	12	1	5.14 (4)	4.31 (4)	1.186×10^{-5} (5)
rs4766443	12	0.474	0.586	3.21 (19)	0.032
rs10774613	12	0.783	0.804 (17)	3.049	6.221×10^{-3} (16)
rs1265566	12	0.601	0.867	3.332 (15)	0.021
rs17696736	12	1	8.774 (3)	4.876 (3)	2.532×10^{-6} (4)
rs12924729	16	0.98	3.46 (5)	3.859 (6)	4.771×10^{-4} (8)
rs7193144	16	0.803	0.747 (18)	3.011	4.761×10^{-3} (14)
rs8050136	16	0.768	0.689 (19)	2.986	5.919×10^{-3} (15)
rs9926289	16	0.519	0.337	2.8	0.018 (20)
rs2542151	18	0.998	2.619 (10)	3.797 (7)	7.514×10^{-5} (7)

SNP details		Summary statistics	
SNP	Chr	p-val (X_3)	FDR (X_4)
rs17013326	1	9.501×10^{-7}	0.026
rs1230666	1	9.121×10^{-10}	8.1×10^{-5}
rs6679677	1	3.583×10^{-14}	2.658×10^{-10}
rs6661817	1	7.411×10^{-6}	0.128
rs3811019	1	3.046×10^{-6}	0.057
rs12061474	1	1.366×10^{-5}	0.234
rs903228	2	0.372	≥ 0.5

rs7666328	4	0.021	≥ 0.5
rs2544677	5	3.278×10^{-6}	0.059
rs2112168	5	9.241×10^{-6}	0.192
rs7917983	10	9.129×10^{-7}	5.483×10^{-3}
rs7901275	10	9.996×10^{-7}	5.325×10^{-3}
rs7901695	10	4.229×10^{-7}	5.471×10^{-10}
rs12243326	10	5.464×10^{-6}	3.365×10^{-7}
rs3741939	12	0.066	0.096
rs705698	12	7.058×10^{-6}	0.016
rs705702	12	5.177×10^{-6}	0.013
rs2292239	12	8.463×10^{-8}	6.252×10^{-5}
rs4766443	12	1.264×10^{-5}	0.208
rs10774613	12	3.07×10^{-5}	0.042
rs1265566	12	6.733×10^{-6}	0.126
rs17696736	12	6.675×10^{-9}	1.263×10^{-5}
rs12924729	16	5.72×10^{-7}	3.029×10^{-3}
rs7193144	16	3.735×10^{-5}	0.033
rs8050136	16	4.257×10^{-5}	0.041
rs9926289	16	1.203×10^{-4}	0.109
rs2542151	18	7.456×10^{-7}	4.5×10^{-4}

Table D.9 Top 20 SNPs differentiating T1D and T2D (MHC removed), considered as subgroups of a general diabetic phenotype, for each of four summary statistics. Positions are in NCBI build 36. Because of the large number of SNPs with evidence for differentiating the subgroups, only SNPs with non-zero weights after applying the LDAK procedure are included in this table. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0'||Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores	
SNP	Chr	Pos	Gene	$\lvert Z_d \rvert (p)$	$\lvert Z_a \rvert$
rs10858002	1	113794974	*PTPN22*	2.997 (2.7×10^{-3})	3.44
rs17013326	1	113801358	*PTPN22*	2.465 (0.01)	4.223
rs1230666	1	113885452	*PTPN22*	5.326 (1×10^{-7})	5.801
rs6679677	1	114015850	*PTPN22*	7.137 (9.5×10^{-13})	6.9
rs3811019	1	114183625	*PTPN22*	2.943 (3.3×10^{-3})	3.132
rs10931347	2	189007813		3.014 (2.6×10^{-3})	2.984
rs6846031	4	178394297		3.474 (5.1×10^{-4})	2.908
rs11970411	6	138220854	*TNFAIP3*	3.601 (3.2×10^{-4})	2.402
rs3114834	7	109192112		3.092 (2×10^{-3})	3.204
rs16874205	8	107271324		3.042 (2.4×10^{-3})	3.866
rs2104286	10	6139051	*IL2RA*	3.911 (9.2×10^{-5})	2.842
rs7917983	10	114722872	*TCF7L2*	2.296 (0.02)	3.823
rs7901275	10	114722896	*TCF7L2*	2.064 (0.04)	4.016
rs7901695	10	114744078	*TCF7L2*	5.689 (1.3×10^{-8})	2.203
rs12243326	10	114778805	*TCF7L2*	4.592 (4.4×10^{-6})	2.087
rs10736277	10	121705898	*TCF7L2*	2.992 (2.8×10^{-3})	3.064
rs770738	12	10034164	*DEXI*	2.364 (0.02)	3.698
rs1495377	12	69863368	*TSPAN8*	4.302 (1.7×10^{-5})	2.063
rs7961581	12	69949369	*TSPAN8*	3.715 (2×10^{-4})	2.935
rs551714	13	20436464		2.381 (0.02)	3.659
rs1711029	15	51491702		2.742 (6.1×10^{-3})	4.722
rs1054028	16	22834715		3.063 (2.2×10^{-3})	3.222
rs7193144	16	52368187	*FTO*	2.469 (0.01)	3.622
rs8050136	16	52373776	*FTO*	2.575 (0.01)	3.548
rs896136	17	35904973	*IKZF3*	3.02 (2.5×10^{-3})	3.122

SNP details		Values (rank)			
SNP	Chr	X_1	X_2	X_3	X_4
rs10858002	1	0.413	0.472 (11)	3.171 (11)	0.042 (11)
rs17013326	1	0.743	0.933 (6)	3.072 (15)	0.029 (8)
rs1230666	1	1	10.802 (2)	5.515 (2)	1.508×10^{-7} (2)
rs6679677	1	1	19.721 (1)	7.039 (1)	9.563×10^{-13} (1)

rs3811019	1	0.207	0.199	3.019 (19)	0.11
rs10931347	2	0.17	0.163	3.002 (20)	0.14
rs6846031	4	0.324	0.379 (13)	3.23 (9)	0.053 (12)
rs11970411	6	0.161	0.181	3.052 (17)	0.162
rs3114834	7	0.305	0.328 (14)	3.137 (12)	0.07 (14)
rs16874205	8	0.748	1.213 (5)	3.355 (7)	0.013 (6)
rs2104286	10	0.554	0.804 (8)	3.433 (4)	0.03 (9)
rs7917983	10	0.356	0.283 (16)	2.828	0.08 (16)
rs7901275	10	0.407	0.268 (17)	2.709	0.093
rs7901695	10	0.987	4.345 (3)	3.861 (3)	6.625×10^{-5} (3)
rs12243326	10	0.587	0.912 (7)	3.326 (8)	0.01 (4)
rs10736277	10	0.194	0.188	3.021 (18)	0.113
rs770738	12	0.298	0.236	2.838	0.084 (19)
rs1495377	12	0.369	0.483 (10)	3.186 (10)	0.032 (10)
rs7961581	12	0.486	0.655 (9)	3.373 (6)	0.025 (7)
rs551714	13	0.28	0.221	2.838	0.086 (20)
rs1711029	15	0.971	2.145 (4)	3.424 (5)	0.011 (5)
rs1054028	16	0.302	0.322 (15)	3.127 (13)	0.074 (15)
rs7193144	16	0.29	0.245	2.888	0.081 (18)
rs8050136	16	0.287	0.255	2.936	0.081 (17)
rs896136	17	0.23	0.231	3.061 (16)	0.094

SNP details		Summary statistics	
SNP	Chr	p-val (X_3)	FDR (X_4)
rs10858002	1	1.492×10^{-5}	0.27
rs17013326	1	2.656×10^{-5}	0.172
rs1230666	1	6.511×10^{-11}	1.391×10^{-6}
rs6679677	1	3.084×10^{-15}	8.924×10^{-12}
rs3811019	1	3.64×10^{-5}	≥ 0.5
rs10931347	2	3.952×10^{-5}	≥ 0.5
rs6846031	4	1.041×10^{-5}	0.326
rs11970411	6	2.948×10^{-5}	≥ 0.5
rs3114834	7	1.822×10^{-5}	0.452
rs16874205	8	5.053×10^{-6}	0.077
rs2104286	10	3.19×10^{-6}	0.185

rs7917983	10	1.079×10^{-4}	0.478
rs7901275	10	2.092×10^{-4}	≥ 0.5
rs7901695	10	3.159×10^{-7}	5.521×10^{-4}
rs12243326	10	6.013×10^{-6}	0.056
rs10736277	10	3.562×10^{-5}	≥ 0.5
rs770738	12	1.021×10^{-4}	0.478
rs1495377	12	1.37×10^{-5}	0.197
rs7961581	12	4.576×10^{-6}	0.145
rs551714	13	1.021×10^{-4}	0.49
rs1711029	15	3.406×10^{-6}	0.058
rs1054028	16	1.903×10^{-5}	0.458
rs7193144	16	7.761×10^{-5}	0.491
rs8050136	16	5.855×10^{-5}	0.489
rs896136	17	2.826×10^{-5}	≥ 0.5

Table D.12 Top 20 SNPs differentiating T2D and RA (MHC removed), considered as subgroups of a general phenotype, for each of four summary statistics. Positions are in NCBI build 36. Because of the large number of SNPs with evidence for differentiating the subgroups, only SNPs with non-zero weights after applying the LDAK procedure are included in this table. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0' \big| |Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0' | X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores	
SNP	Chr	Pos	Gene	$\lvert Z_d \rvert (p)$	$\lvert Z_a \rvert$
rs6679677	1	114105331	*PTPN22*	2.568 (0.01)	9.84
rs2476601	1	114179091	*PTPN22*	2.649 (8.1×10^{-3})	9.88
rs7554023	1	160162988	*ATF6??*	3.625 (2.9×10^{-4})	1.855
X2-204400444-CA-DELETION	2	204400444	*CTLA4*	2.142 (0.03)	8.822
X2-204408002-CCT-DELETION	2	204408002	*CTLA4*	2.103 (0.04)	8.879
rs58716662	2	204423821	*CTLA4*	2.447 (0.01)	5.968
rs78960870	2	204458162	*CTLA4*	2.171 (0.03)	6.143
rs13030124	2	204402508	*CTLA4*	2.091 (0.04)	8.871
rs3997876	2	179005067	*PRKRA*	7.863 (3.8×10^{-15})	10.102
rs3997878	2	179004872	*PRKRA*	7.358 (1.9×10^{-13})	9.467
rs6720771	2	154461782		4.091 (4.3×10^{-5})	0.428
rs6723546	2	154617139		4.137 (3.5×10^{-5})	0.76
rs12638263	3	187278549	*BCL6*	3.459 (5.4×10^{-4})	2.263
rs34244025	9	138290559	*ESP33*	4.649 (3.3×10^{-6})	1.426
rs34775390	9	138293196	*ESP33*	4.595 (4.3×10^{-6})	1.508
rs6582394	12	40972456		3.247 (1.2×10^{-3})	2.504
rs10220315	14	80197971	*CEP128*	2.947 (3.2×10^{-3})	4.818
rs10136185	14	80210225	*CEP128*	2.954 (3.1×10^{-3})	4.579
rs78304225	14	80276765	*CEP128*	2.853 (4.3×10^{-3})	4.249
rs327443	14	80291769	*CEP128*	3.331 (8.7×10^{-4})	6.025
rs327465	14	80299793	*CEP128*	4.731 (2.2×10^{-6})	6.653
rs55957493	14	80539807	*CEP128*	5.634 (1.8×10^{-8})	9.916
rs17545310	14	80540892	*CEP128*	2.844 (4.5×10^{-3})	6.223
rs2284734	14	80623486	*CEP128*	2.81 (5×10^{-3})	4.358
rs2284735	14	80623539	*CEP128*	2.99 (2.8×10^{-3})	3.781

SNP details		Values (rank)			
SNP	Chr	X_1	X_2	X_3	X_4
rs6679677	1	1	1.994 (8)	4.019 (7)	0.01 (14)
rs2476601	1	1	2.224 (7)	4.109 (5)	8.063×10^{-3} (11)

rs7554023	1	0.11	0.064	2.899	0.012 (18)
X2-204400444-CA-DELETION	2	1	1.015 (11)	3.435 (10)	0.044
X2-204408002-CCT-DELETION	2	1	0.907 (12)	3.399 (12)	0.041
rs58716662	2	1	1.93 (9)	3.294 (14)	0.02
rs78960870	2	1	1.335 (10)	3.071 (20)	0.043
rs13030124	2	1	0.879 (13)	3.385 (13)	0.042
rs3997876	2	1	25.336 (1)	8.548 (1)	3.008×10^{-14} (1)
rs3997878	2	1	22.077 (2)	8.003 (2)	1.777×10^{-12} (2)
rs6720771	2	0.05	0.051	1.927	0.015 (20)
rs6723546	2	0.07	0.065	2.352	8.97×10^{-3} (12)
rs12638263	3	0.176	0.116	3.003	0.011 (17)
rs34244025	9	0.414	0.497	3.135 (19)	6.905×10^{-4} (6)
rs34775390	9	0.414	0.494	3.169 (18)	6.897×10^{-4} (5)
rs6582394	12	0.199	0.126	2.977	0.014 (19)
rs10220315	14	0.999	2.934	3.472 (9)	5.334×10^{-3} (8)
rs10136185	14	0.996	2.85	3.419 (11)	5.943×10^{-3} (9)
rs78304225	14	0.977	2.341	3.258 (15)	0.011 (16)
rs327443	14	1	4.141 (5)	4.059 (6)	1.628×10^{-3} (7)
rs327465	14	1	8.77 (4)	5.301 (4)	7.38×10^{-6} (4)
rs55957493	14	1	13.509 (3)	6.803 (3)	2.346×10^{-8} (3)
rs17545310	14	1	2.882 (6)	3.692 (8)	7.092×10^{-3} (10)
rs2284734	14	0.984	2.366	3.253 (16)	0.011 (15)
rs2284735	14	0.899	1.764	3.234 (17)	9.448×10^{-3} (13)

SNP details		Summary statistics	
SNP	Chr	p-val (X_3)	FDR (X_4)
rs6679677	1	2.433×10^{-6}	0.045
rs2476601	1	1.784×10^{-6}	0.035
rs7554023	1	1.529×10^{-4}	0.054
X2-204400444-CA-DELETION	2	1.803×10^{-5}	0.193
X2-204408002-CCT-DELETION	2	2.047×10^{-5}	0.179

rs58716662	2	3.026×10^{-5}	0.089
rs78960870	2	7.274×10^{-5}	0.188
rs13030124	2	2.146×10^{-5}	0.184
rs3997876	2	8.23×10^{-17}	1.046×10^{-13}
rs3997878	2	5.231×10^{-15}	6.2×10^{-12}
rs6720771	2	0.011	0.065
rs6723546	2	1.833×10^{-3}	0.039
rs12638263	3	9.761×10^{-5}	0.05
rs34244025	9	5.625×10^{-5}	2.932×10^{-3}
rs34775390	9	4.904×10^{-5}	2.923×10^{-3}
rs6582394	12	1.093×10^{-4}	0.061
rs10220315	14	1.576×10^{-5}	0.024
rs10136185	14	1.912×10^{-5}	0.027
rs78304225	14	3.453×10^{-5}	0.047
rs327443	14	2.122×10^{-6}	7.152×10^{-3}
rs327465	14	1.677×10^{-8}	2.329×10^{-5}
rs55957493	14	8.75×10^{-12}	7.965×10^{-8}
rs17545310	14	7.401×10^{-6}	0.031
rs2284734	14	3.53×10^{-5}	0.047
rs2284735	14	3.782×10^{-5}	0.041

Table D.15 Top 20 SNPs differentiating Graves' disease and Hashimoto's thyroiditis (MHC removed), considered as subgroups of autoimmune thyroid disease, for each of four summary statistics. Positions are in NCBI build 37. Because of the density of the genotyping chip used and the large number of SNPs with evidence of differentiating the subgroups, with non-zero weights after applying the LDAK procedure are included in this table. The column 'LDAK' gives the weight attributed to the SNP by the LDAK procedure. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0'||Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores	
SNP	Chr	Pos	Gene	$\lvert Z_d \rvert (p)$	$\lvert Z_a \rvert$
rs231790	2	204408819	*CTLA4*	3.465 (5.3×10^{-4})	5.584
rs231797	2	204414352	*CTLA4*	3.466 (5.3×10^{-4})	5.567
rs231804	2	204416891	*CTLA4*	3.046 (2.3×10^{-3})	6.22
rs11571304	2	204417021	*CTLA4*	3.047 (2.3×10^{-3})	6.22
rs3087243	2	204447164	*CTLA4*	3.355 (7.9×10^{-4})	6.606
rs6748358	2	204465150	*CTLA4*	3.395 (6.9×10^{-4})	5.887
rs7596727	2	204491827	*CTLA4*	3.578 (3.5×10^{-4})	5.065
rs2352551	2	204503002	*CTLA4*	3.558 (3.7×10^{-4})	5.247
rs3757247	6	91014184	*BACH2*	4.037 (5.4×10^{-5})	3.683
rs11755527	6	91014952	*BACH2*	4.105 (4×10^{-5})	3.936
rs619192	6	91025670	*BACH2*	3.135 (1.7×10^{-3})	4.791
rs1847472	6	91029880	*BACH2*	3.465 (5.3×10^{-4})	4.389
rs604912	6	91043041	*BACH2*	3.144 (1.7×10^{-3})	4.762
rs17251453	12	91026211		4.26 (2×10^{-5})	2.593
rs12426486	12	91052228		4.433 (9.3×10^{-6})	2.671
rs7334298	13	41359622		4.442 (8.9×10^{-6})	2.566
rs9525555	13	41408305		4.44 (9×10^{-6})	2.525
rs9532960	13	41443676		4.377 (1.2×10^{-5})	2.418
rs16967120	15	36707739	*RASGRP1*	3.088 (2×10^{-3})	4.359
rs2839511	21	42721590	*UBASH3A*	3.533 (4.1×10^{-4})	4.47

SNP details		Values (rank)			
SNP	Chr	X_1	X_2	X_3	X_4
rs231790	2	0.979	2.672 (6)	3.825	0.047
rs231797	2	0.979	2.677 (5)	3.824	0.047
rs231804	2	0.925	1.764	3.531	0.031 (6)
rs11571304	2	0.925	1.764	3.532	0.032 (7)
rs3087243	2	0.94	2.163 (10)	3.86 (7)	0.036
rs6748358	2	0.969	2.471 (9)	3.805	0.044
rs7596727	2	0.987	2.958 (2)	3.845 (9)	0.044
rs2352551	2	0.986	2.904 (3)	3.856 (8)	0.043
rs3757247	6	0.952	2.611 (7)	3.961 (4)	0.015 (2)

rs11755527	6	0.98	3.265 (1)	4.069 (1)	0.015 (1)
rs619192	6	0.971	2.124	3.423	0.03 (5)
rs1847472	6	0.975	2.579 (8)	3.638	0.02 (3)
rs604912	6	0.971	2.137	3.426	0.033 (9)
rs17251453	12	0.674	0.831	3.844 (10)	0.071
rs12426486	12	0.788	1.171	3.992 (2)	0.053
rs7334298	13	0.752	1.013	3.965 (3)	0.066
rs9525555	13	0.735	0.948	3.951 (5)	0.037
rs9532960	13	0.657	0.727	3.871 (6)	0.033 (8)
rs16967120	15	0.949	1.896	3.317	0.035 (10)
rs2839511	21	0.981	2.76 (4)	3.709	0.029 (4)

SNP details		Summary statistics	
SNP	Chr	p-val (X_3)	FDR (X_4)
rs231790	2	2.179×10^{-6}	0.067
rs231797	2	2.204×10^{-6}	0.069
rs231804	2	1.07×10^{-5}	0.034
rs11571304	2	1.064×10^{-5}	0.036
rs3087243	2	1.812×10^{-6}	0.044
rs6748358	2	2.45×10^{-6}	0.062
rs7596727	2	1.946×10^{-6}	0.062
rs2352551	2	1.845×10^{-6}	0.059
rs3757247	6	1.046×10^{-6}	0.017
rs11755527	6	5.814×10^{-7}	0.016
rs619192	6	1.913×10^{-5}	0.032
rs1847472	6	5.968×10^{-6}	0.024
rs604912	6	1.866×10^{-5}	0.039
rs17251453	12	1.951×10^{-6}	0.12
rs12426486	12	8.797×10^{-7}	0.079
rs7334298	13	1.017×10^{-6}	0.107
rs9525555	13	1.102×10^{-6}	0.048
rs9532960	13	1.696×10^{-6}	0.038
rs16967120	15	3.361×10^{-5}	0.043
rs2839511	21	4.072×10^{-6}	0.03

Table D.18 Top ten SNPs differentiating TPOA positive and negative T1D (MHC removed), for each of four summary statistics. Positions are in NCBI build 36. Only SNPs with positive weights after applying the LDAK procedure (and therefore used in fitting the model) were considered here. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0'||Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

SNP details				Z scores	
SNP	Chr	Pos	Gene	$\lvert Z_d \rvert (p)$	$\lvert Z_a \rvert$
rs231790	2	204408819	*CTLA4*	2.815 (4.9×10^{-3})	5.584
rs231797	2	204414352	*CTLA4*	2.795 (5.2×10^{-3})	5.567
rs11571293	2	204425958	*CTLA4*	2.834 (4.6×10^{-3})	5.995
rs6748358	2	204465150	*CTLA4*	2.847 (4.4×10^{-3})	5.887
rs7596727	2	204491827	*CTLA4*	2.944 (3.2×10^{-3})	5.065
rs2352551	2	204503002	*CTLA4*	2.949 (3.2×10^{-3})	5.247
rs1560418	3	159972335		4.541 (5.6×10^{-6})	1.21
rs1560417	3	159972476		4.543 (5.5×10^{-6})	1.213
rs511198	4	116234541		4.685 (2.8×10^{-6})	1.858
rs506851	4	116234970		4.746 (2.1×10^{-6})	1.956
rs503256	4	116244902		4.645 (3.4×10^{-6})	1.832
rs473989	4	116246844		4.687 (2.8×10^{-6})	1.781
rs505277	4	116248257		4.634 (3.6×10^{-6})	1.81
rs1507935	4	116368809		4.586 (4.5×10^{-6})	1.514
rs867036	4	116381578		4.581 (4.6×10^{-6})	1.515
rs7694946	4	116413588		4.693 (2.7×10^{-6})	1.963
rs706781	10	6126391	*IL2RA*	3.098 (1.9×10^{-3})	4.515
rs907092	17	35175785	*IKZF3*	3.746 (1.8×10^{-4})	4.221
rs11078927	17	35317931	*IKZF3*	3.774 (1.6×10^{-4})	4.168
rs4795400	17	35320546	*IKZF3*	3.569 (3.6×10^{-4})	4.106

SNP details		Values (rank)			
SNP	Chr	X_1	X_2	X_3	X_4
rs231790	2	0.971	1.953 (7)	2.978	0.143
rs231797	2	0.971	1.924 (8)	2.957	0.117
rs11571293	2	0.966	2.044 (4)	3.013	0.271
rs6748358	2	0.968	2.048 (3)	3.022	0.282
rs7596727	2	0.978	1.977 (6)	3.078	0.204
rs2352551	2	0.978	2.042 (5)	3.091	0.364
rs1560418	3	5.957×10^{-3}	0.11	4.075 (10)	0.082
rs1560417	3	5.988×10^{-3}	0.11	4.078 (9)	0.09
rs511198	4	0.025	0.131	4.344 (3)	0.039 (4)

rs506851	4	0.033	0.141	4.413 (1)	0.07 (10)
rs503256	4	0.023	0.128	4.305 (5)	0.037 (3)
rs473989	4	0.021	0.128	4.33 (4)	0.045 (7)
rs505277	4	0.022	0.127	4.29 (6)	0.033 (1)
rs1507935	4	0.011	0.116	4.188 (7)	0.055 (8)
rs867036	4	0.011	0.115	4.184 (8)	0.056 (9)
rs7694946	4	0.032	0.138	4.37 (2)	0.088
rs706781	10	0.964	1.862 (10)	3.195	0.401
rs907092	17	0.958	2.227 (1)	3.783	0.045 (6)
rs11078927	17	0.951	2.175 (2)	3.805	0.045 (5)
rs4795400	17	0.927	1.868 (9)	3.61	0.035 (2)

SNP details		Summary statistics	
SNP	Chr	p-val (X_3)	FDR (X_4)
rs231790	2	7.2×10^{-4}	0.146
rs231797	2	7.831×10^{-4}	0.117
rs11571293	2	6.162×10^{-4}	0.447
rs6748358	2	5.989×10^{-4}	0.494
rs7596727	2	4.648×10^{-4}	0.276
rs2352551	2	4.349×10^{-4}	≥ 0.5
rs1560418	3	3.168×10^{-6}	0.275
rs1560417	3	3.108×10^{-6}	0.088
rs511198	4	6.607×10^{-7}	0.137
rs506851	4	4.33×10^{-7}	0.227
rs503256	4	8.292×10^{-7}	0.131
rs473989	4	7.184×10^{-7}	0.166
rs505277	4	9.08×10^{-7}	0.113
rs1507935	4	1.644×10^{-6}	0.168
rs867036	4	1.709×10^{-6}	0.172
rs7694946	4	5.593×10^{-7}	0.096
rs706781	10	2.764×10^{-4}	≥ 0.5
rs907092	17	1.565×10^{-5}	0.163
rs11078927	17	1.392×10^{-5}	0.161
rs4795400	17	3.785×10^{-5}	0.121

Table D.21 Top ten SNPs with differing effect sizes with age at diagnosis in T1D (MHC removed), for each of four summary statistics. Positions are in NCBI build 36. Only SNPs with positive weights after applying the LDAK procedure (and therefore used in fitting the model) were considered here. Ranks in X_2 (bracketed) are only amongst SNPs with $X_1 > 0.7$; ranks in X_3 and X_4 are amongst all SNPs. The value X_1 is the posterior probability of category 3 membership (SNPs differentiating subgroups); X_2 is the contribution to the pseudo-likelihood ratio from the SNP; X_3 is a weighted geometric mean of Z_a and Z_d and X_4 is the conditional false discovery rate for observations z_a and z_d at the SNP; that is, $Pr(H_0'||Z_d| \leq |z_d|, |Z_a| \leq |z_a|)$, where H_0' is the hypothesis that the SNP has the same population minor allele frequencies in subgroups. P-values are computed based on X_3, under the null hypothesis that (Z_a, Z_d) has a joint mixture bivariate Gaussian distribution consistent with H_0. A value $X_4 = \alpha$ does not correspond to a false-discovery rate of α amongst SNPs with $X_4 \leq \alpha$; the corresponding value, $P(H_0'|X_4 < \alpha)$ is given in the rightmost column. Potential gene associations are marked.

D.2 Supplementary figures

Please see following page

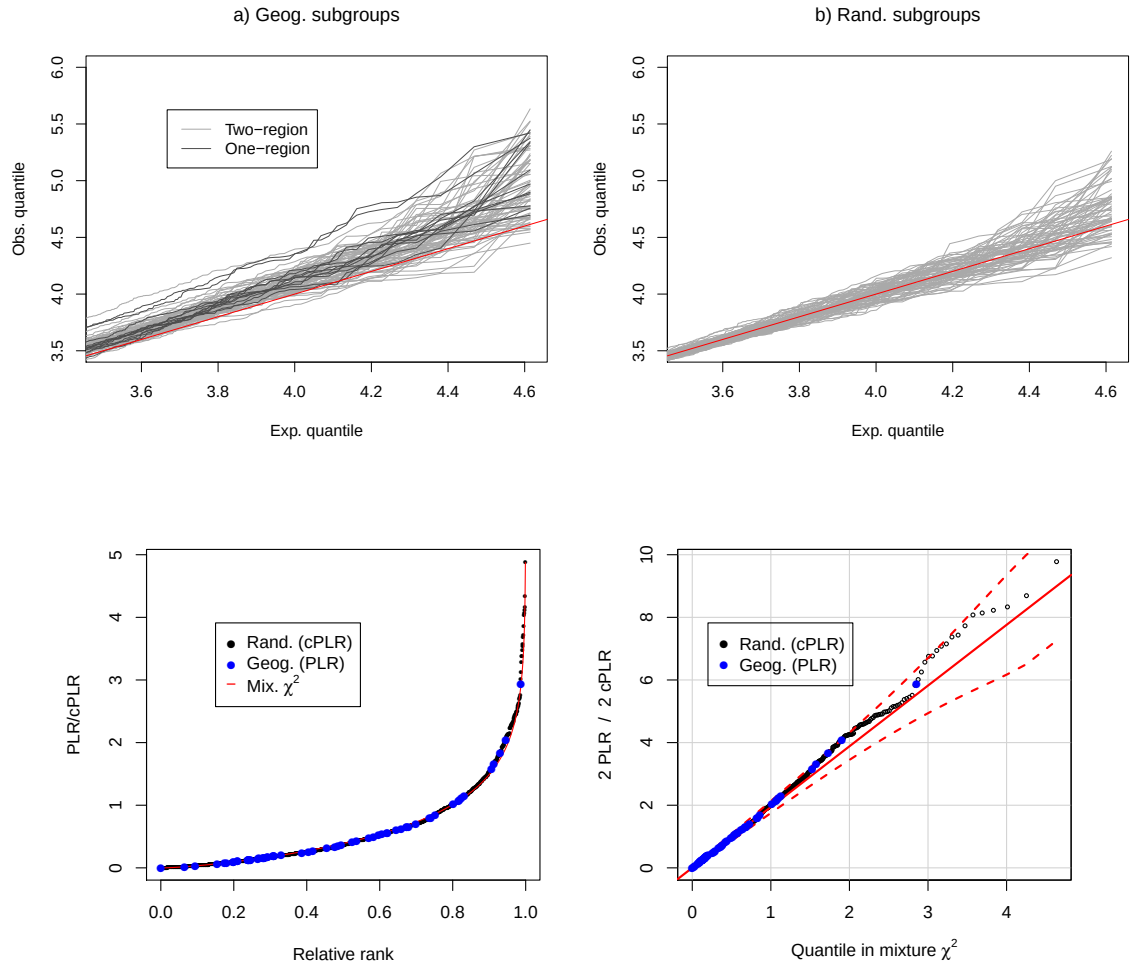


Fig. D.1 Top plot shows Z_d scores arising from geography-based subgroups compared with expected normal. Leftmost plot shows quantiles of Z scores from geography based subgroups; two-region subgroups in light grey and one-region subgroups in dark grey. Considerable inflation is seen compared to Z-scores arising from random subgroups, in rightmost figure. Lower plots show distribution of cPLR values from random subgroups against observed PLR values from geographically-defined subgroups. Leftmost plot shows cPLR values from random subgroups plotted in ascending with PLR values from random subgroups shown in blue. Rightmost plot is Q-Q plot comparing null cPLR distribution with the asymptotic mixture-χ^2.

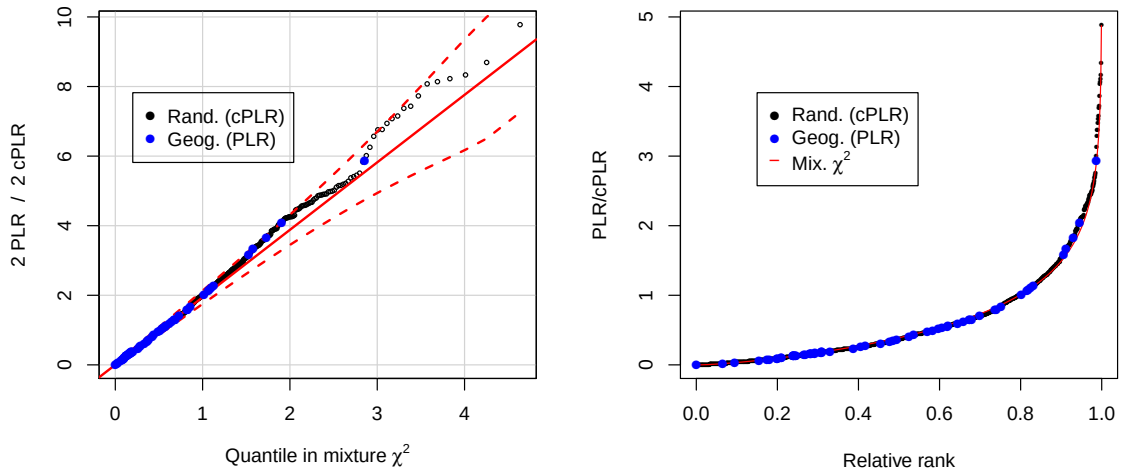
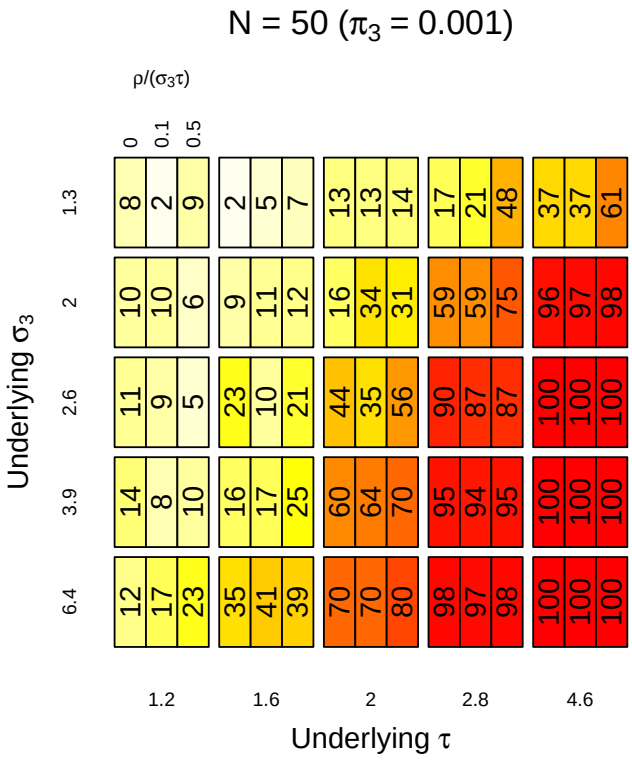
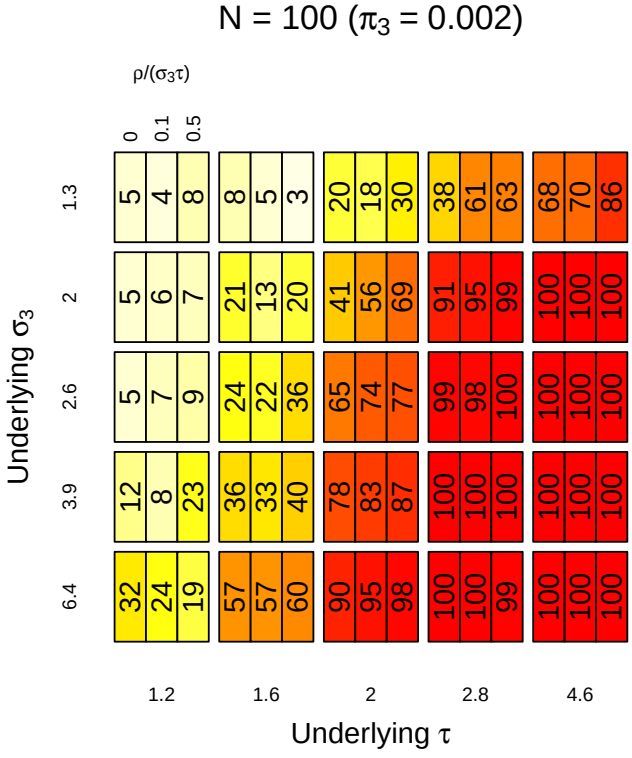
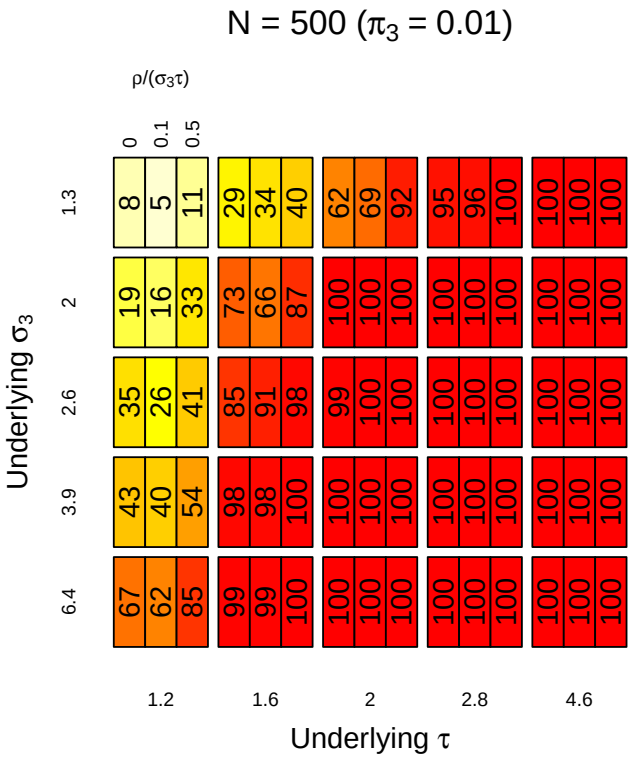
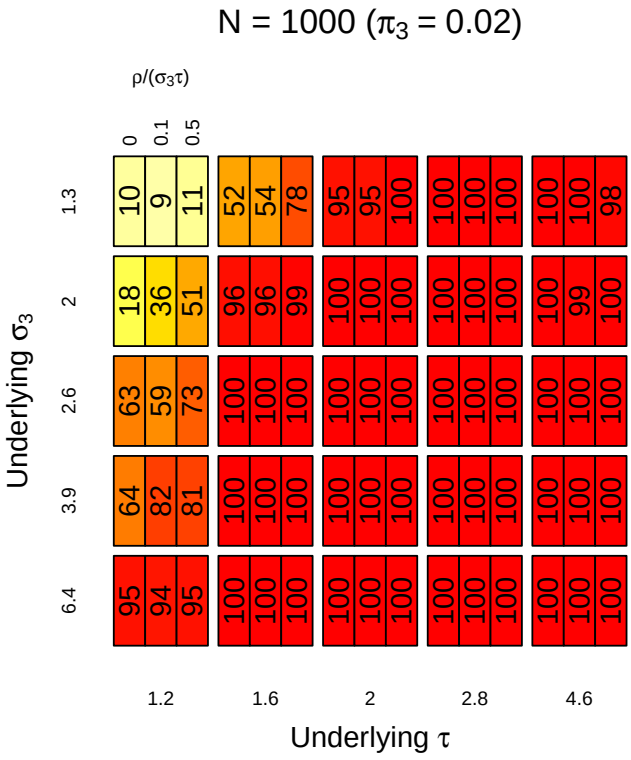


Fig. D.2 Summary of test statistics (PLR) from geographically-defined subgroups, based on WTCCC data [The Wellcome Trust Case Control Consortium, 2007] for controls and type 1 diabetes (T1D). In each instance, one subgroup was defined as the controls coming from either one or two geographic regions, and the other subgroup as the controls coming from the remaining nine or ten geographic regions. I also generated > 2000 randomly allocated subgroups and computed the cPLR. The left panel shows a Q-Q plot of cPLR values from random subgroups against the asymptotic mixture-χ^2 distribution, with blue points representing the PLRs of geographic subgroups. The right panel shows cPLR values plotted in ascending order with the PLR values from geographic subgroups included as blue points. The minimum Bonferroni-corrected empirical p value was > 0.5


N = 50 (π3 = 0.001)
ρ/(σ3τ)
0
0.1
0.5
Underlying σ3
1.3
2
2.6
3.9
6.4
8 2 9 | 2 5 7 | 13 13 14 | 17 21 48 | 37 37 61
10 10 6 | 9 11 12 | 16 34 31 | 59 59 75 | 96 97 98
11 9 5 | 23 10 21 | 44 35 56 | 90 87 87 | 100 100 100
14 8 10 | 16 17 25 | 60 64 70 | 95 94 95 | 100 100 100
12 17 23 | 35 41 39 | 70 70 80 | 98 97 98 | 100 100 100
1.2
1.6
2
2.8
4.6
Underlying τ


N = 100 (π3 = 0.002)
ρ/(σ3τ)
0
0.1
0.5
Underlying σ3
1.3
2
2.6
3.9
6.4
5 4 8 | 8 5 3 | 20 18 30 | 38 61 63 | 68 70 86
5 6 7 | 21 13 20 | 41 56 69 | 91 95 99 | 100 100 100
5 7 9 | 24 22 36 | 65 74 77 | 99 98 100 | 100 100 100
12 8 23 | 36 33 40 | 78 83 87 | 100 100 100 | 100 100 100
32 24 19 | 57 57 60 | 90 95 98 | 100 100 99 | 100 100 100
1.2
1.6
2
2.8
4.6
Underlying τ


N = 500 (π_3 = 0.01)
ρ/(σ3τ)
0 0.1 0.5
Underlying σ3
1.3 8 5 11 29 34 40 62 69 92 95 96 100 100 100
2 19 16 33 73 66 87 100 100 100 100 100 100 100 100 100
2.6 35 26 41 85 91 98 99 100 100 100 100 100 100 100 100
3.9 43 40 54 98 98 100 100 100 100 100 100 100 100 100 100
6.4 67 62 85 99 99 100 100 100 100 100 100 100 100 100 100
1.2 1.6 2 2.8 4.6
Underlying τ


N = 1000 (π_3 = 0.02)
ρ/(σ3τ)
0 0.1 0.5
Underlying σ3
1.3 10 9 11 52 54 78 95 95 100 100 100 100 100 100 98
2 18 36 51 96 96 99 100 100 100 100 100 100 100 99 100
2.6 63 59 73 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
3.9 64 82 81 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
6.4 95 94 95 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
1.2 1.6 2 2.8 4.6
Underlying τ

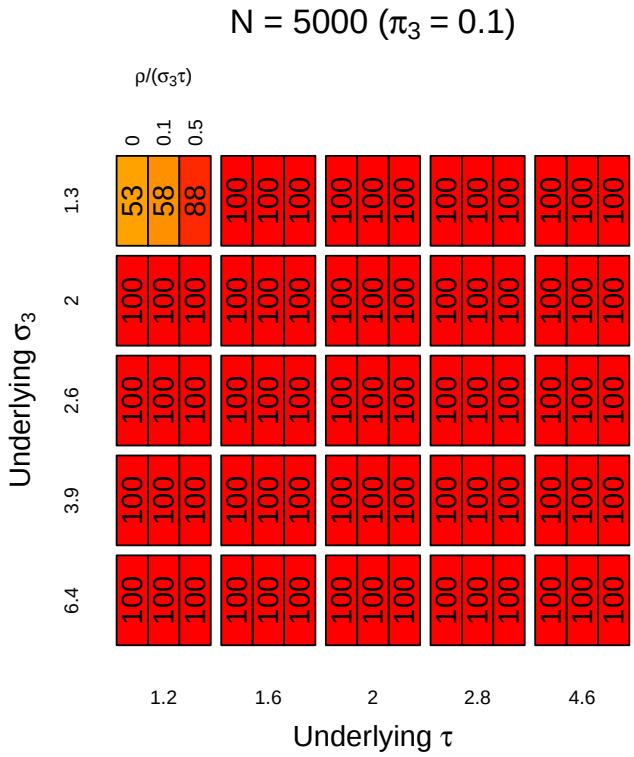
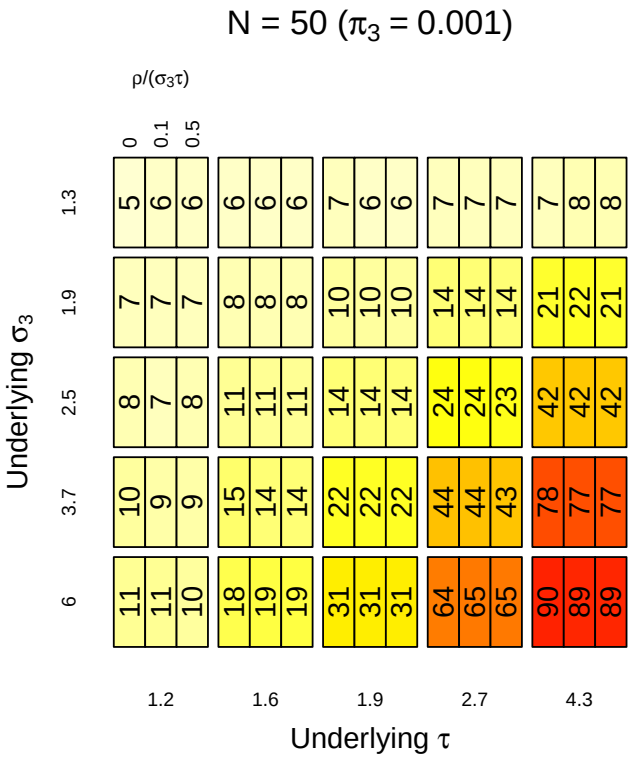
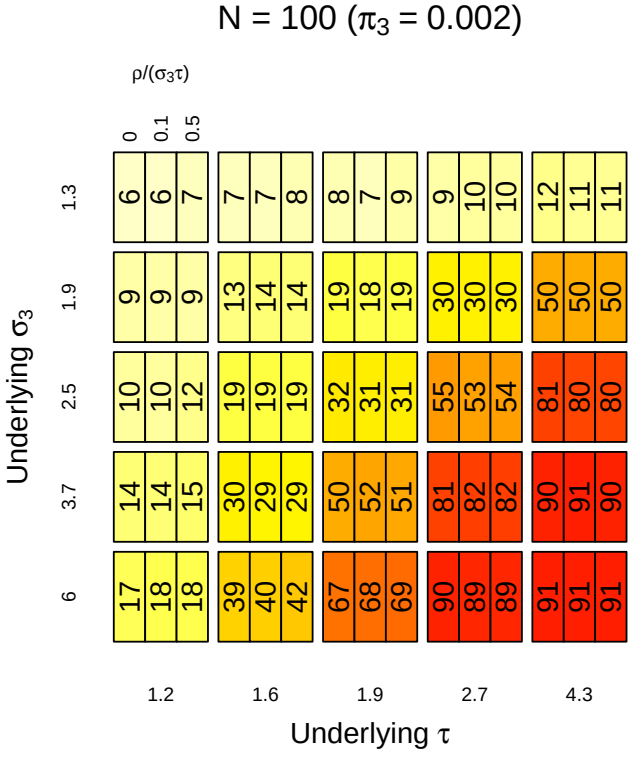
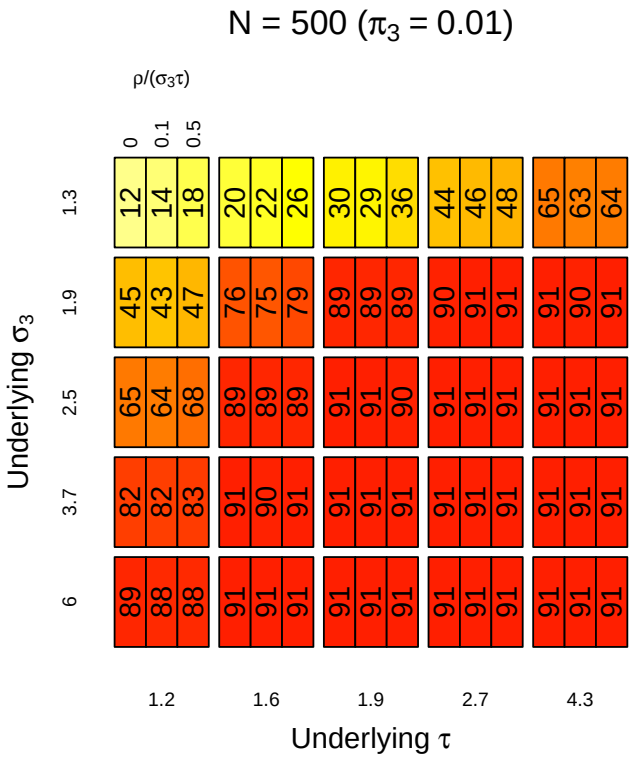
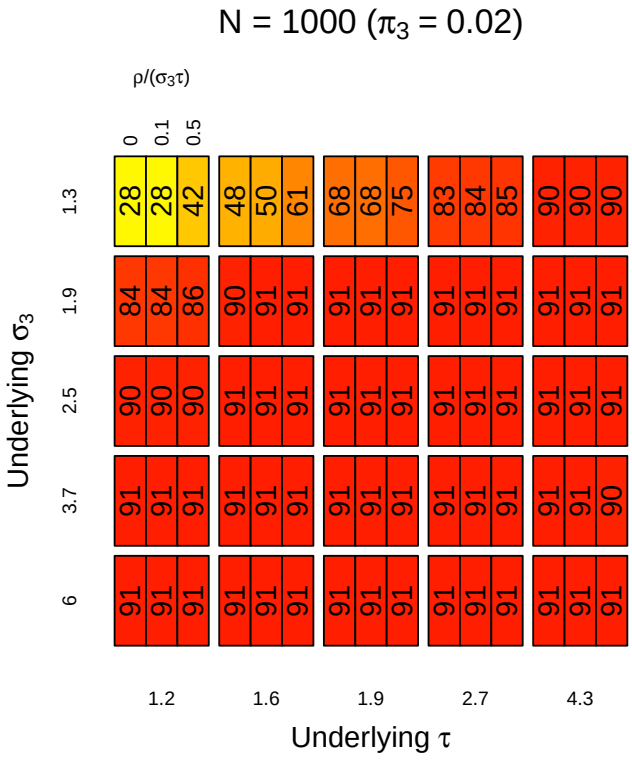


Fig. D.3 Estimates of power to reject H_0 for subgrouping problem ($\alpha = 0.05$) for various values of π_3, σ_3, τ, and ρ. The value N is the approximate number of SNPs in category 3, corresponding to π_3. In total, each simulation was on 5×10^4 simulated autosomal SNPs in linkage equilibrium. The value $\rho/(\sigma_3\tau)$ is the correlation (rather than covariance) between Z_a and Z_d in category 3.


N = 50 (π_3 = 0.001)
ρ/(σ3τ)
0 0.1 0.5
Underlying σ3
1.3 1.9 2.5 3.7 6
5 6 6 | 6 6 6 | 7 6 6 | 7 7 7 | 7 8 8
7 7 7 | 8 8 8 | 10 10 10 | 14 14 14 | 21 22 21
8 7 8 | 11 11 11 | 14 14 14 | 24 24 23 | 42 42 42
10 9 9 | 15 14 14 | 22 22 22 | 44 44 43 | 78 77 77
11 11 10 | 18 19 19 | 31 31 31 | 64 65 65 | 90 89 89
1.2 1.6 1.9 2.7 4.3
Underlying τ


N = 100 (π_3 = 0.002)
ρ/(σ3τ)
0 0.1 0.5
Underlying σ3
1.3 1.9 2.5 3.7 6
6 6 7 | 7 7 8 | 8 7 9 | 9 10 10 | 12 11 11
9 9 9 | 13 14 14 | 19 18 19 | 30 30 30 | 50 50 50
10 10 12 | 19 19 19 | 32 31 31 | 55 53 54 | 81 80 80
14 14 15 | 30 29 29 | 50 52 51 | 81 82 82 | 90 91 90
17 18 18 | 39 40 42 | 67 68 69 | 90 89 89 | 91 91 91
1.2 1.6 1.9 2.7 4.3
Underlying τ


N = 500 (π_3 = 0.01)
ρ/(σ3τ)
0
0.1
0.5
Underlying σ3
1.3
1.9
2.5
3.7
6
12 14 18
20 22 26
30 29 36
44 46 48
65 63 64
45 43 47
76 75 79
89 89 89
90 91 91
91 90 91
65 64 68
89 89 89
91 91 90
91 91 91
91 91 91
82 82 83
91 90 91
91 91 91
91 91 91
91 91 91
89 88 88
91 91 91
91 91 91
91 91 91
91 91 91
1.2
1.6
1.9
2.7
4.3
Underlying τ


N = 1000 (π_3 = 0.02)
ρ/(σ3τ)
0
0.1
0.5
Underlying σ3
1.3
1.9
2.5
3.7
6
28 28 42
48 50 61
68 68 75
83 84 85
90 90 90
84 84 86
90 91 91
91 91 91
91 91 91
91 91 91
90 90 90
91 91 91
91 91 91
91 91 91
91 91 91
91 91 91
91 91 91
91 91 91
91 91 91
91 91 90
91 91 91
91 91 91
91 91 91
91 91 91
91 91 91
1.2
1.6
1.9
2.7
4.3
Underlying τ

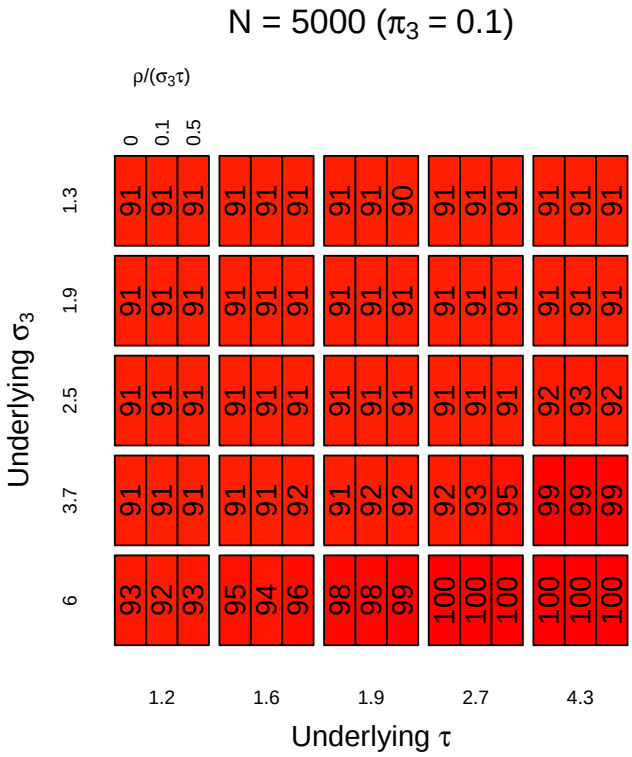
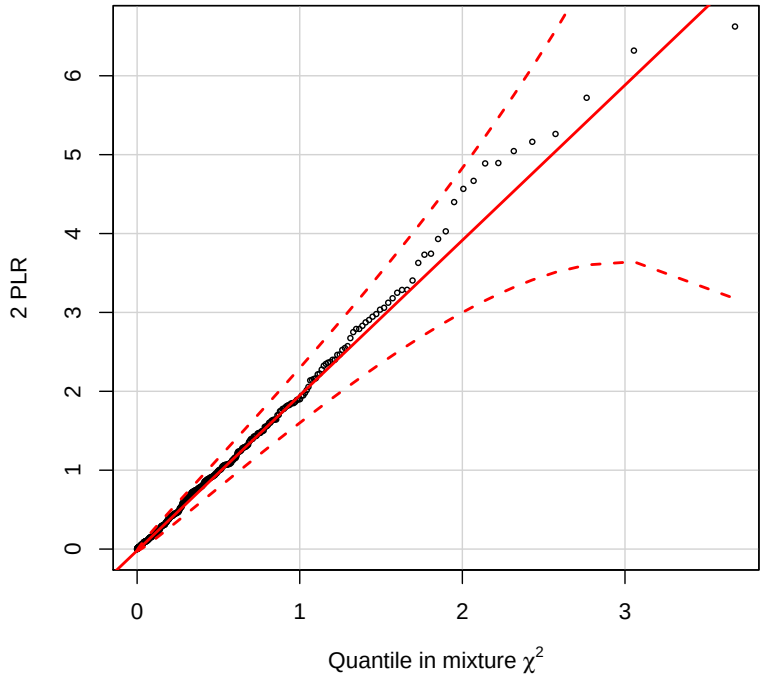
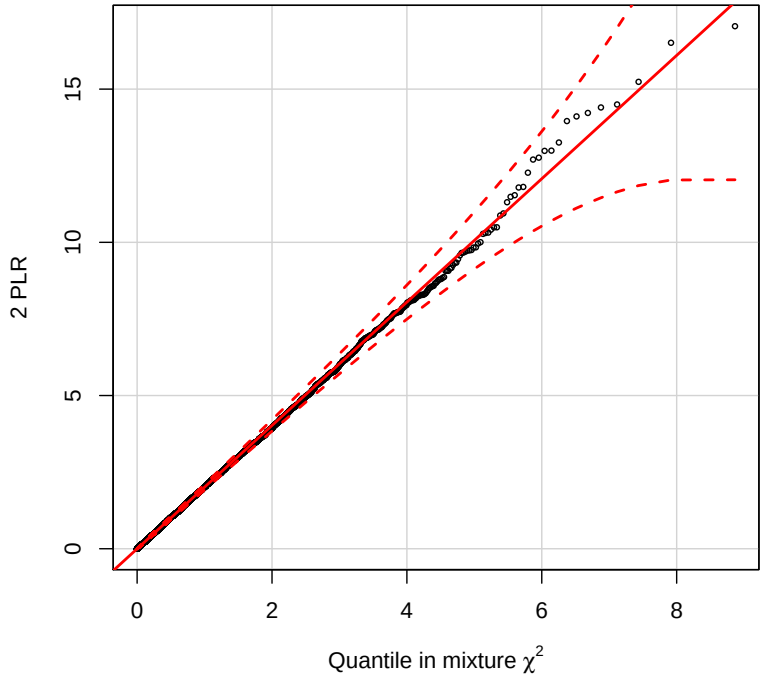


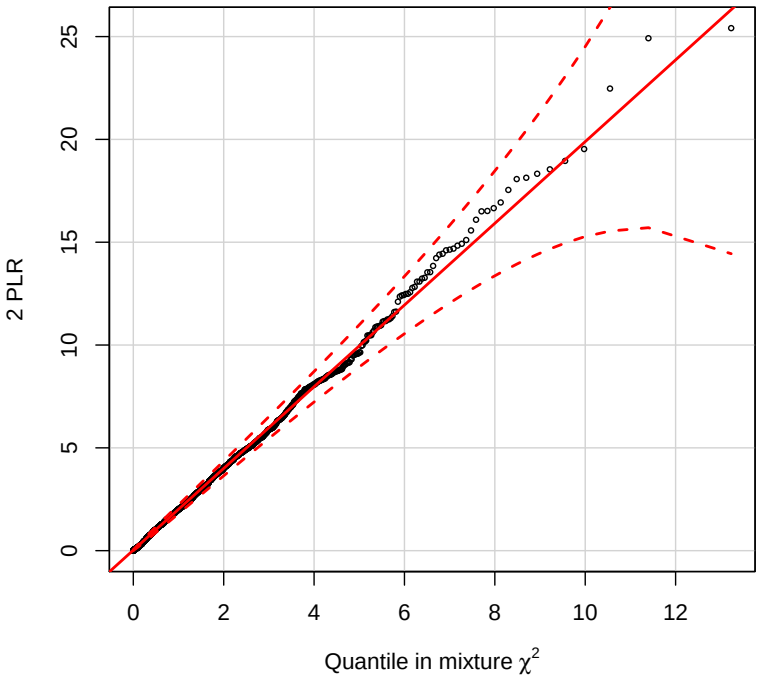
Fig. D.4 Estimates of power of alternative test statistic (K) to reject H_0 ($\alpha = 0.05$) for subgrouping problem for various values of π_3, σ_3, τ, and ρ. The value N is the approximate number of SNPs in category 3, corresponding to π_3. In total, each simulation was on 5×10^4 simulated autosomal SNPs in linkage equilibrium. The value $\rho/(\sigma_3\tau)$ is the correlation (rather than covariance) between Z_a and Z_d in category 3.



(a) T1D/T2D/RA data

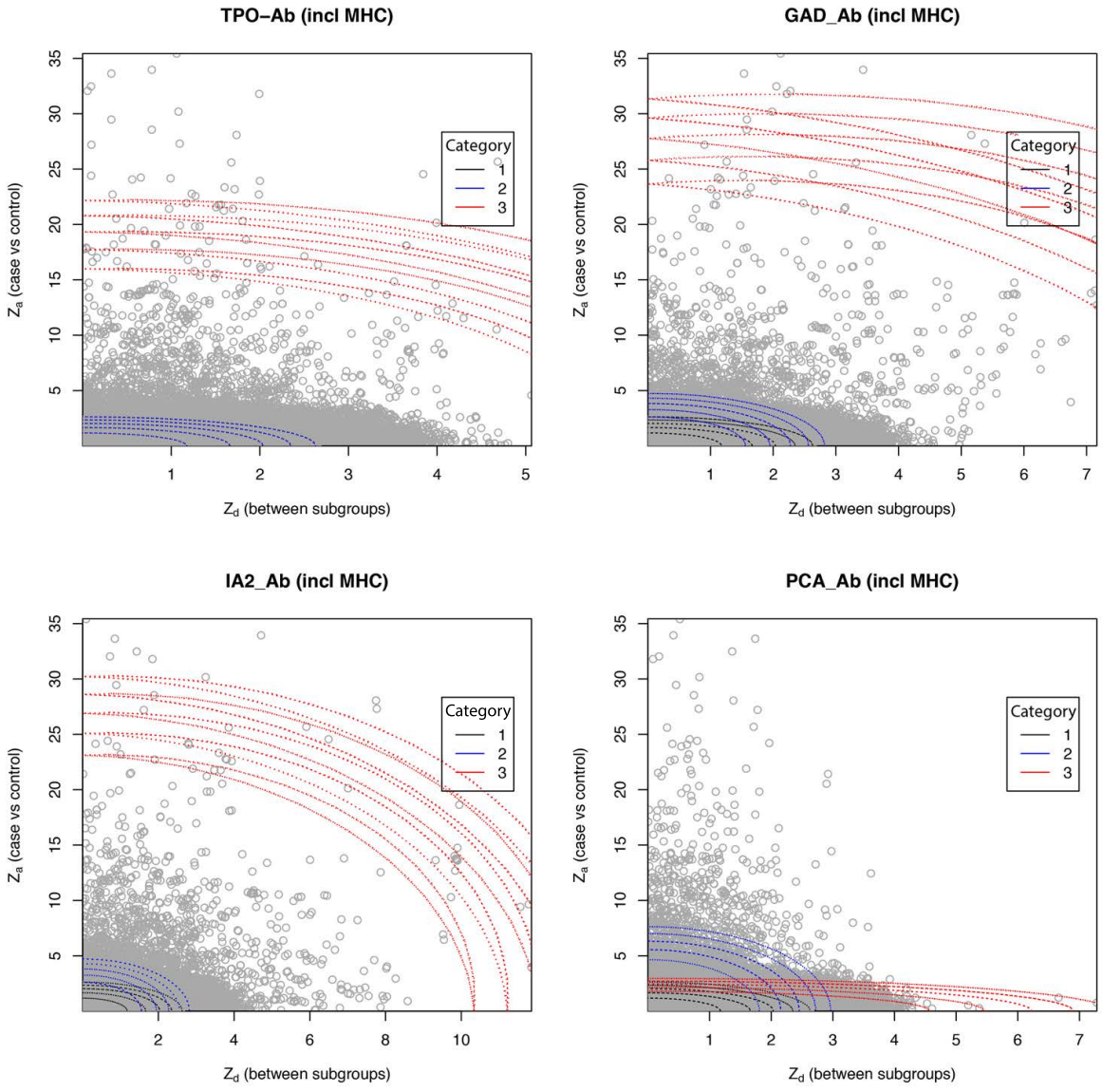


(b) ATD:GH/HT data



(c) T1D (AAB) data

Fig. D.5 Q-Q plot of the distribution of observed test statistics (cPLR) for random subgroups of tested phenotypes (T1D/RA/T2D combined, GH/HT combined, T1D) against a mixture χ^2 distribution of the form $\gamma * (\kappa\chi_1^2 + (1-\kappa)\chi_2^2)$. A 99% confidence interval is shown by the dashed red lines. The distribution is well-approximated by the asymptotic mixture-χ^2 in all cases.


TPO−Ab (incl MHC)
GAD_Ab (incl MHC)
IA2_Ab (incl MHC)
PCA_Ab (incl MHC)
Category
1
2
3
Z_a (case vs control)
Z_d (between subgroups)

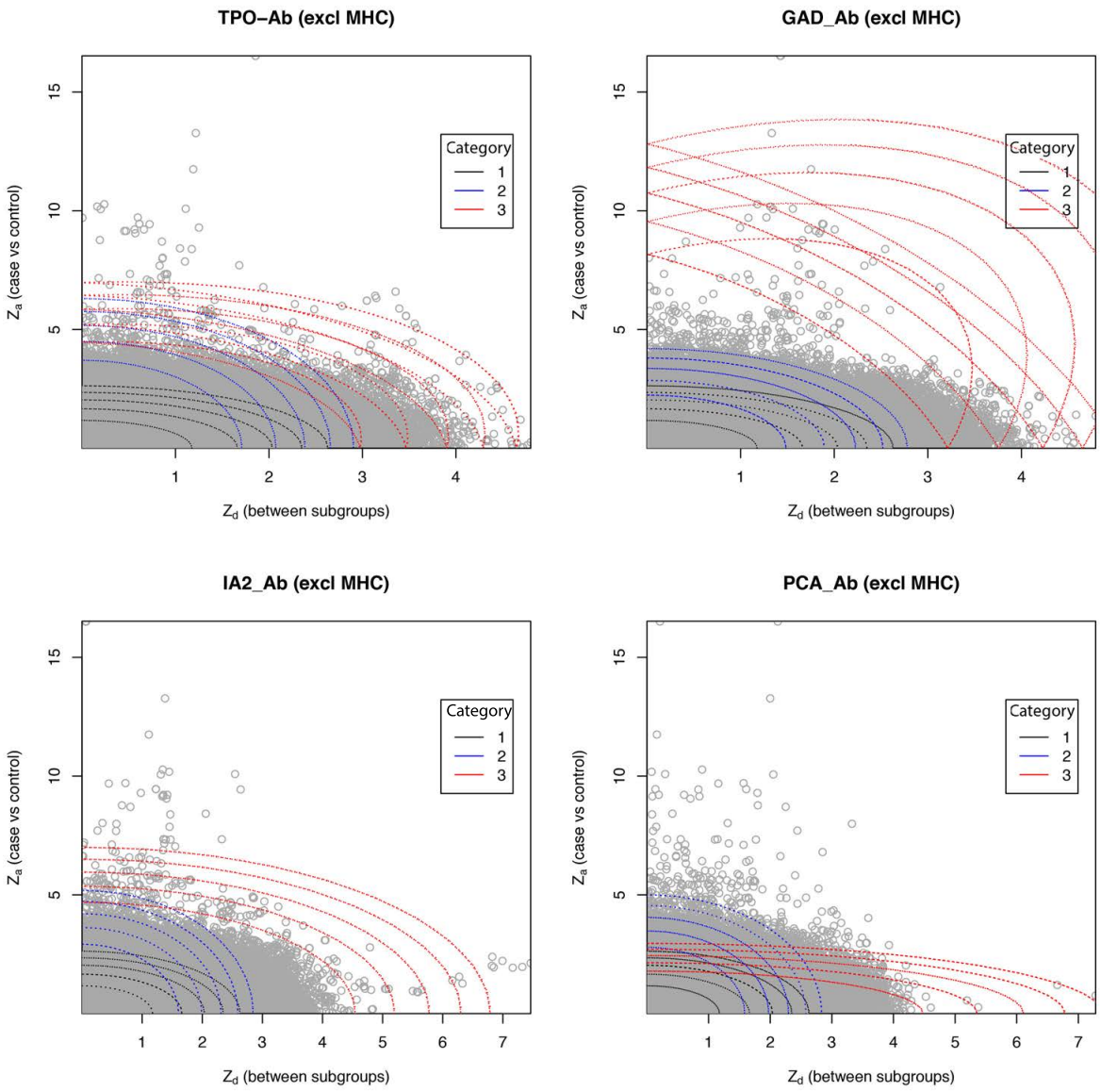


Fig. D.6 Observed Z_a and Z_d scores (grey) for T1D subtypings based on autoantibody positivity, including or excluding the MHC region, and contours of parameters of fitted models (coloured ellipses). Full models are shown for the comparisons involving TPO-Ab, GAD-Ab, and IA2-Ab, and null models for PCA-Ab (for which the null hypothesis could not be rejected). Note the differing X-axis scales. The plots illustrate the rationale for the three-category model; for TPO-Ab, GAD-Ab and IA2-Ab, a tendency is seen for SNPs associated with autoantibody positivity (high $|Z_d|$) to be associated with T1D also (high $|Z_a|$). This tendency is not seen for PCA-Ab, and is minimal for non-MHC SNPs in GAD-Ab. Further analysis of the plot for TPOAb positivity (top left) is shown below.

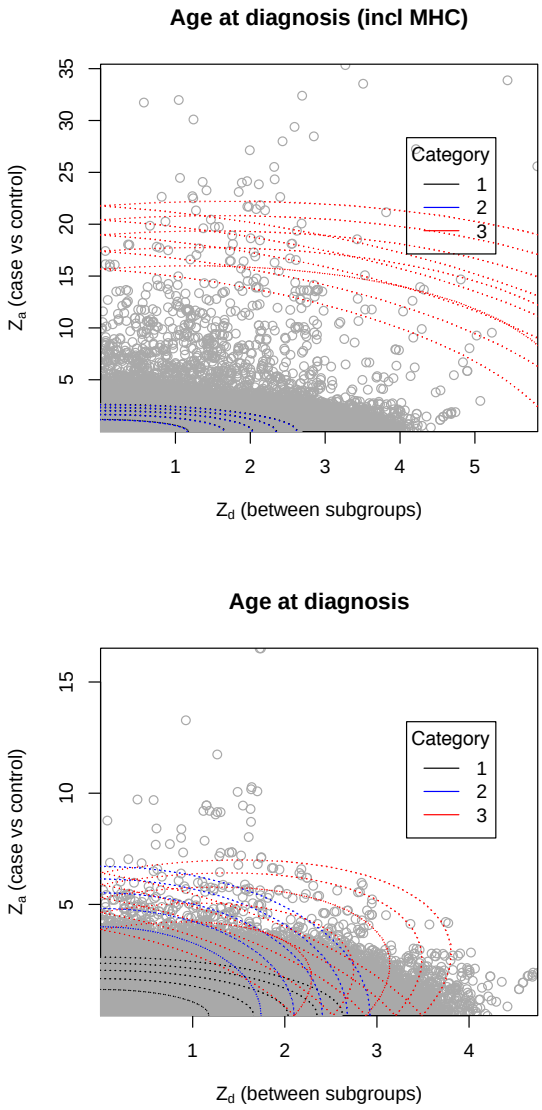
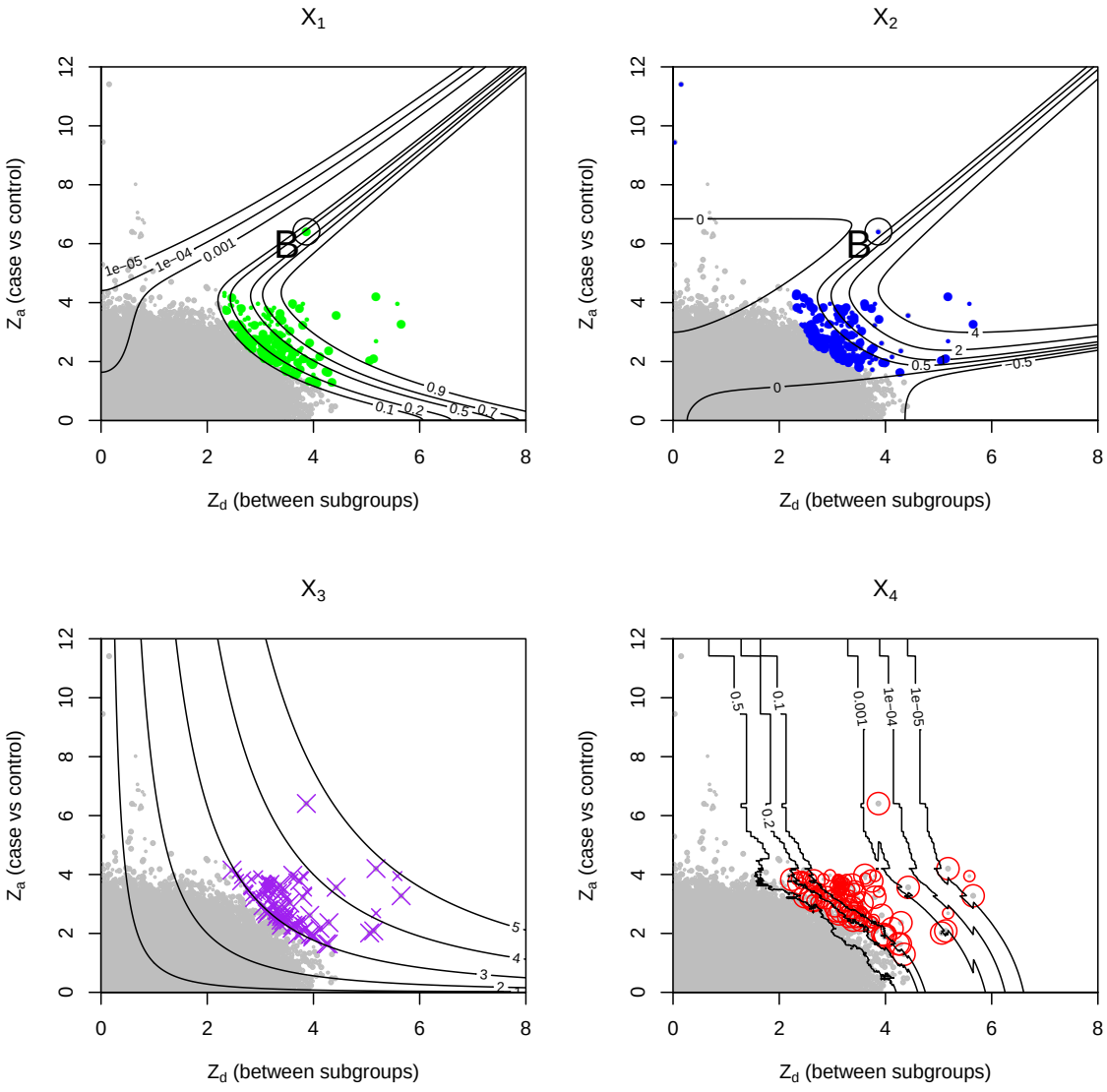
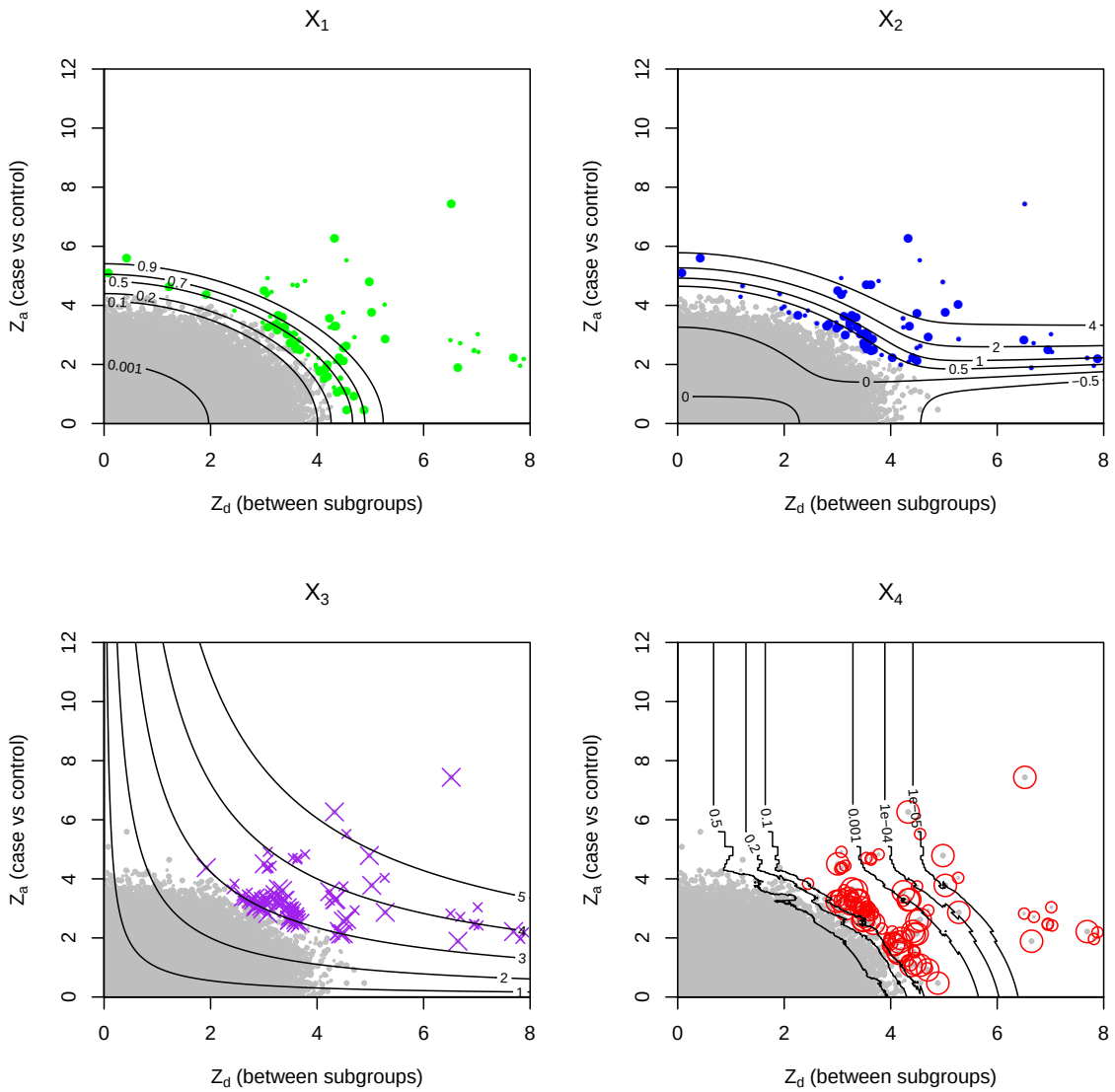


Fig. D.7 Observed Z_a and Z_d scores for T1D subclassified by age at diagnosis. Non-MHC SNPs are shown in red.

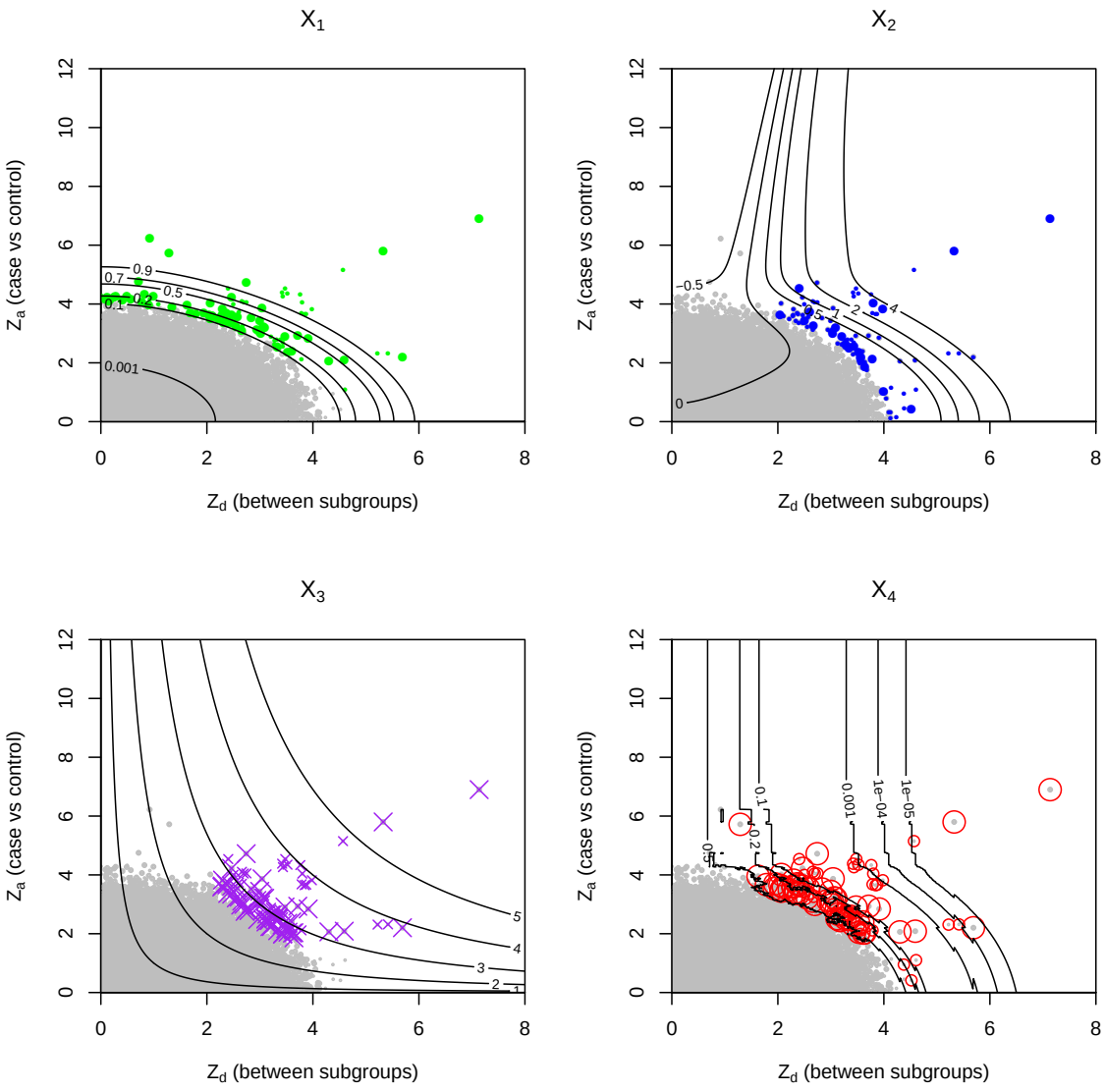
(a) T1D/RA comparison



(b) T1D/T2D comparison



(c) T2D/RA comparison



(d) GD/HT (ATD) comparison

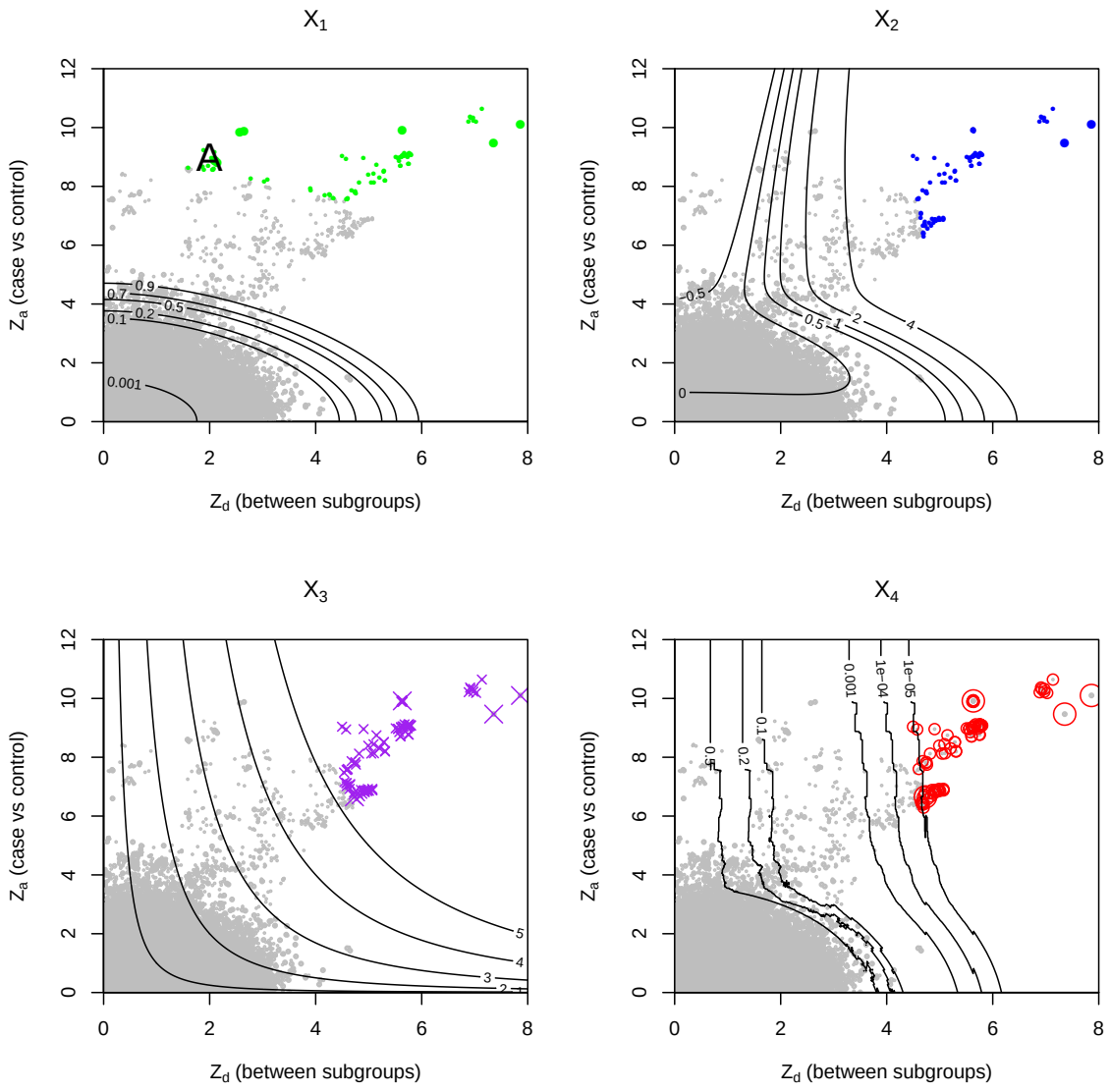


Fig. D.8 I demonstrate all four test statistics for single-SNP effects in the comparisons betwen T1D/T2D/RA, and between GD and HT (preceding pages). The top 100 SNPs for each test statistic are highlighted, with larger symbols corresponding to SNPs with non-zero weights after applying LDAK [Speed et al., 2012]; that is, the SNPs which contributed to the model fit. Contours of each test statistic are shown in grey.

Differences are evident in the behaviour of the test statistics X_1 and X_2 between the two datasets; X_3 and X_4 are more robust. The different null hypotheses between X_3 and X_4 are responsible for the difference in shape near the line $Z_a = 0$. Contours of X_4 are jagged due to the dependence of this statistic on the distribution of Z scores.

All methods primarily identified SNPs with both high $|Z_a|$ and $|Z_d|$ scores as contributors. As evident from the comparison between GH and HT, the statistic X_1 is vulnerable to falsely declaring SNPs as subgroup-differentiating despite low $|Z_d|$ scores (labelled 'A', top left panel, GD/HT). This arises due to the full model having a markedly higher value of σ_3 than σ_2, leading to SNPs with very high $|Z_a|$ values having a high posterior probability of category 3 membership.

This is partially able to be overcome by combining the test statistics X_1 and X_2 into one, which I typically do by only considering X_2 scores in SNPs with X_1 greater than some cutoff. However, this is not always effective, as is evident from the above figure for T1D/T2D. In this case, as discussed in the main paper, almost all SNPs with high Z_a also had high Z_d, meaning that the two distributions forming categories 2 and 3 under the null model were essentially the same. This led to the fitted parameters of the null model supporting SNPs falling into two distributions; one with identity covariance matrix, and the other with $\text{var}(Z_d) > 1$, $\text{var}(Z_a) = 1$ (see fitted parameters).

The different alternative hypothesis for X_4 (different population MAFs in subgroups without requiring association with the phenotype overall) meant that SNPs with low $|Z_a|$ scores may be identified by X_4 in addition to those identified by X_1, X_2 and X_3 (contour lines on bottom right panel, both figures). SNPs which are isolated may be missed by both X_1 and X_2 (label 'B', top two panels, T1D/RA), due to the fitted distribution of SNPs in category 3 tending to be driven by clusters of SNPs.

Given these results, I consider X_3 and X_4 to generally be the most appropriate measure for single SNP effects, although in appropriate circumstances X_2 can be used alone or conditionally on X_1.

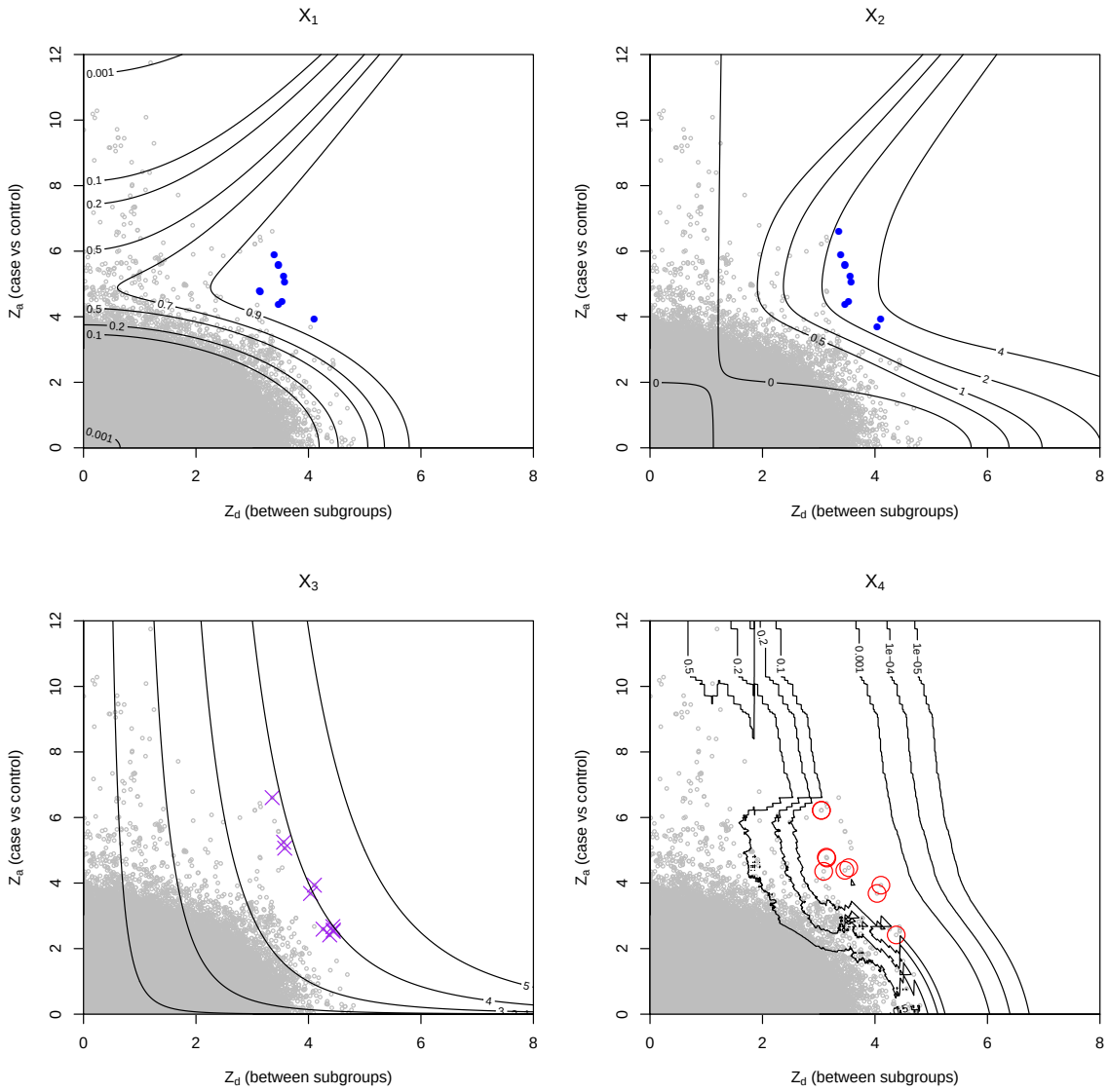


Fig. D.9 I assessed the SNPs responsible for the observed difference in pseudo-likelihood ratio for my analysis of TPOAb positivity in T1D. SNPs in the MHC region were removed from the analysis (co-ordinates 25-38 Mb, GChR build 37). I combined X_1 and X_2 into a single test statistic, by only considering SNPs with $X_1 > 0.7$ and then considering the top SNPs for X_2. The top ten SNPs for $X_2|X_1 > 0.7$ (blue, top two panels), X_3 (purple, bottom left panel), and X_4 (red, bottom right panel) are shown. Contours of each summary statistic are shown as black lines. Details of SNPs are shown in appendix D.1.

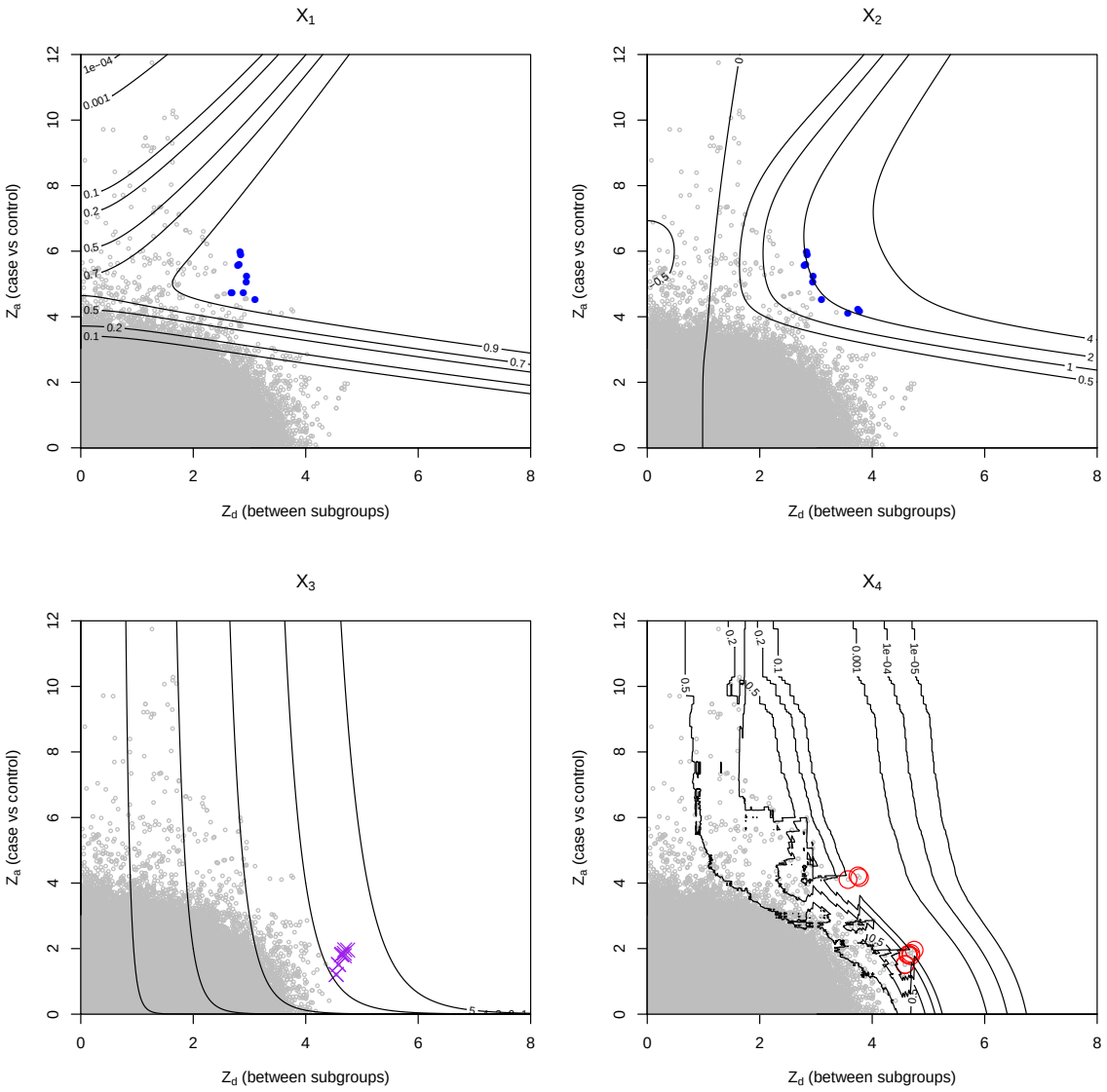


Fig. D.10 I assessed the SNPs responsible for the observed difference in pseudo-likelihood ratio for my analysis of age at diagnosis in T1D. SNPs in the MHC region were removed from the analysis (co-ordinates 25-38 Mb, GChR build 37). I combined X_1 and X_2 into a single test statistic, by only considering SNPs with $X_1 > 0.7$ and then considering the top SNPs for X_2. The top ten SNPs for $X_2|X_1 > 0.7$ (blue, top two panels), X_3 (purple, bottom left panel), and X_4 (red, bottom right panel) are shown. Contours of each summary statistic are shown as black lines. Details of SNPs are shown in appendix D.1.